LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

numero 174
febbraio 1983
anno XVI
volume XXX

Migrazione fra «centro» e «periferia»

Per secoli si è avuto nei paesi sviluppati un movimento di popolazione dalle zone periferiche non industrializzate a quelle centrali. Negli anni settanta questa tendenza ha avuto per la prima volta un'inversione

di Daniel R. Vining, Jr.

Sin dall'inizio della Rivoluzione Industriale la transizione da un'economia agricola a un'economia industriale è stata accompagnata in ogni paese da una crescente concentrazione della popolazione. In America Settentrionale, in Europa e in Giappone questa concentrazione è andata avanti ancora per molto tempo dopo che era stato raggiunto un alto livello di sviluppo. Prima del periodo industriale la popolazione che viveva al di fuori delle poche città di una certa importanza era distribuita in base alla produttività del suolo. Via via però che diminuiva la domanda di manodopera agricola e aumentava quella di manodopera industriale, gli abitanti delle zone rurali si trasferirono in gran numero nelle città. Di conseguenza, la distribuzione della popolazione nei paesi industrializzati diventò sempre più irregolare. Le aree urbane, densamente popolate, diventarono qualcosa di diverso dall'«hinterland», scarsamente popolato. Su una scala geografica più vasta, intere regioni diventarono centri d'industria e di commercio, mentre altre regioni conservarono il loro carattere agricolo.

In questo articolo chiamerò centro la zona industriale più importante di un paese, caratterizzata dalla massima densità demografica. Il resto del paese, dove la densità della popolazione è inferiore, sarà indicato col termine periferia. Per fare un esempio, il bacino di Parigi è il centro della Francia; il resto del paese è la periferia. Negli Stati Uniti le tradizionali zone industriali degli stati nordorientali e di quelli medio-occidentali costituiscono il centro, con il sud e l'ovest che fanno da periferia. Nella maggior parte dei casi la zona industriale più importante di un paese sviluppato ha una posizione centrale anche dal punto di vista geografico ed è sempre accessibile alla periferia e ai principali partner commerciali del paese. Il centro comprende le città più grandi e di solito anche la capitale. Date le connessioni tra vitalità economica, crescita demografica e potere politico, il centro spesso predomina all'interno della nazione tanto sul piano politico quanto su quello culturale.

A partire grosso modo dal 1750 e sino alla fine della seconda guerra mondiale, in tutti i paesi nei quali aveva avuto inizio l'industrializzazione il centro si espanse a detrimento della periferia. Dal 1945 però, e soprattutto negli anni settanta, nei paesi economicamente sviluppati si è assistito a una spettacolare inversione del movimento migratorio. In molte delle nazioni più ricche del mondo la direzione della migrazione interna netta non è più dalla periferia al centro. Negli Stati Uniti, il primo paese dove l'inversione ebbe inizio, e in Canada si registra oggi un flusso notevole dal centro alla periferia. Anche nei paesi industriali dell'Europa continentale nordoccidentale il flusso, pur non essendo altrettanto imponente, è diretto verso le zone periferiche. In buona parte del resto dell'Europa occidentale e in Giappone la migrazione verso il centro è scesa a un livello tale da produrre un certo equilibrio tra la migrazione in entrata e quella in uscita dal centro.

Nell'Europa orientale il flusso migratorio è ancora verso il centro. Nel Terzo Mondo, dove lo sviluppo economico è meno avanzato, la migrazione verso le regioni centrali è tuttora notevole. Esiste pertanto attualmente una notevole differenza fra l'andamento migratorio interno dei paesi sviluppati e quello dei paesi in via di sviluppo. Questo mutamento repentino di una tendenza demografica fondamentale non era previsto dagli studiosi di scienze sociali, non solo, ma è tuttora ignoto in larga misura al di fuori del piccolo gruppo di economisti, geografi e demografi che lo stanno studiando. Non è da escludere che la migrazione dalle regioni centrali che ebbe inizio negli anni settanta sia il primo indice di una nuova forma di organizzazione demografica caratterizzata da una distribuzione della popolazione e del potere politico più regolare di quella che prevale attualmente nelle società industriali.

Sono dieci anni ormai che raccolgo dati sulla migrazione interna di 22 paesi, che ho suddiviso in cinque gruppi sulla base dell'andamento migratorio; a conti fatti, questi raggruppamenti sono in larga misura anche geografici. Nel primo gruppo vi sono il Canada e gli Stati Uniti. Nel secondo sono presenti cinque paesi molto industrializzati dell'Europa nordoccidentale: Belgio, Da-

La distribuzione della popolazione in Europa è suggerita da questa fotografia notturna ripresa nel 1977 nell'ambito dell'Air Force Defense Meteorological Satellite Program degli Stati Uniti. Nelle notti serene un satellite registrava la luce proveniente dalla superficie della Terra. Le sorgenti luminose di maggiore rilievo sono le concentrazioni industriali e le aree urbane, che in larga misura coincidono. In questo modo le zone luminose della fotografia corrispondono ad aree geografiche caratterizzate da un'elevata concentrazione sia di industrie sia di popolazione. La corrispondenza tuttavia non è perfetta, in quanto alcune zone densamente popolate hanno un'illuminazione notturna più intensa di altre. In ogni paese la regione da noi definita «centro» comprende la maggior parte delle industrie, buona parte della popolazione e spesso la capitale. Con il termine «periferia» abbiamo invece indicato il resto del paese. La zona luminosa di particolare spicco nella Francia settentrionale è il bacino di Parigi, che è il centro del paese. La zona luminosa a nord e a est di Parigi comprende il centro del Belgio, dell'Olanda e della Germania Occidentale. Il Belgio appare sproporzionatamente luminoso per via dell'intensa illuminazione delle sue autostrade.

nimarca, Francia, Germania Occidentale e Olanda. Il terzo gruppo comprende i paesi esterni dell'Europa occidentale e due paesi della zona del Pacifico: Finlandia, Giappone, Inghilterra, Islanda, Italia, Norvegia, Nuova Zelanda, Spagna e Svezia. Al quarto gruppo appartengono quattro paesi dell'Europa orientale: Cecoslovacchia, Germania Orientale, Polonia e Ungheria. Corea del Sud e Formosa (Taiwan) compongono il gruppo finale.

I 22 paesi del mio campione comprendono molte delle nazioni più ricche ed economicamente più avanzate e una rappresentanza più ridotta di nazioni a un livello di sviluppo economico leggermente inferiore. I dati di cui mi sono servito per calcolare i tassi migratori interregionali sono tratti dai censimenti nazionali e dai registri della popolazione che sono tenuti in quasi tutti i paesi europei e in Giappone, nella Corea del Sud e a Formosa. Questi registri riguardano ogni circoscrizione amministrativa a livello di provincia o di prefettura. Quando si trasferisce in una nuova zona il cittadino è tenuto a segnalare il proprio nuovo indirizzo; nello stesso tempo vengono registrati anche il vecchio indirizzo e la regione di provenienza. Questi registri documentano inoltre i dati statistici demografici riguardanti le nascite e le morti.

Essi permettono quindi di calcolare ogni anno la popolazione di ogni unità amministrativa. La registrazione delle persone che nell'anno sono entrate e uscite dalla zona può essere usata insieme al computo della popolazione, per calcolare il tasso annuo di migrazione fra regione e regione. Unendo fra loro varie zone amministrative, è possibile definire con una certa approssimazione il centro di ogni paese e determinare il tasso di migrazione tra il centro e la periferia. Ovviamente i confini delle circoscrizioni amministrative unite in questo modo non coincidono esattamente con quelli del centro. Con una scelta accurata delle varie zone è possibile per altro definire il centro con una buona approssimazione.

Analizzare la popolazione delle circoscrizioni amministrative ufficiali anziché quella del centro vero e proprio può avere certi vantaggi. È stata avanzata l'ipotesi che nelle nazioni sviluppate l'attuale migrazione dal centro non rappresenti altro che l'espansione del vecchio centro verso le aree immediatamente adiacenti. Scegliendo le giuste zone ufficiali è possibile definire una regione più ampia del centro vero e proprio. Nel caso in cui esista effettivamente, un flusso migratorio netto da questo centro «sovradelimitato» rafforza notevolmente l'ipotesi secondo la quale ciò che sta avvenendo non è soltanto un riversamento in zone limitrofe, ma un allontanamento dal centro esistente.

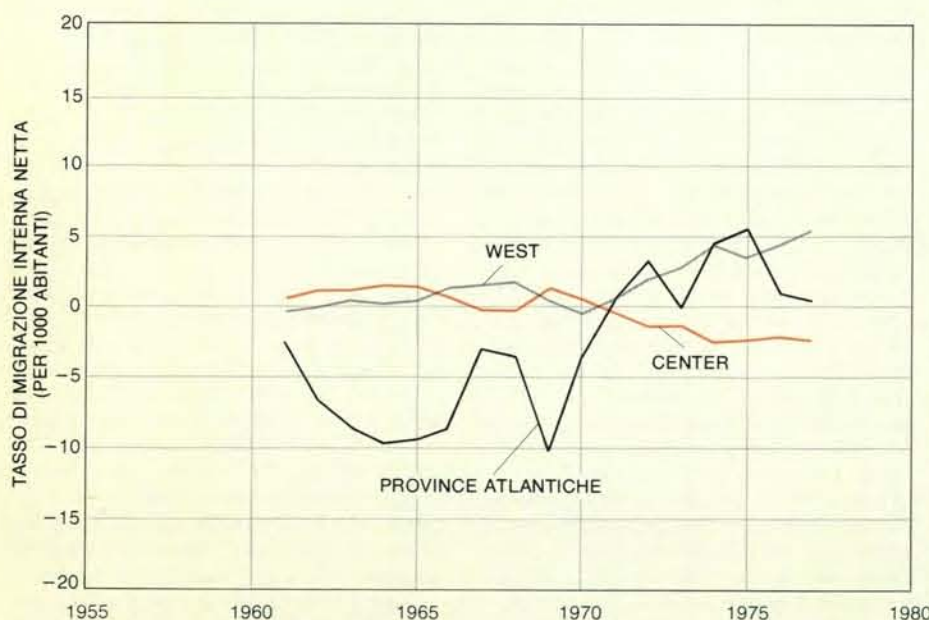
Per capire fino in fondo l'importanza del cambiamento di direzione della migrazione interna che ha avuto luogo negli anni settanta in buona parte del mondo sviluppato è necessario capire perché le popolazioni nazionali si erano concentrate in origine in determinate zone. Per spiegare le ragioni per le quali la crescita dell'industria porta alla concentrazione demografica, gli economisti hanno ipotizzato che nell'organizzazione di una società industriale vi siano delle economie di scala. Un'economia di scala è il risparmio che deriva dal fare un investimento in una zona in cui la scala delle iniziative e la concentrazione delle persone e dell'industria sono già elevate. L'esistenza di tali economie implica che un investimento effettuato nel centro dà profitti più elevati di un investimento della stessa entità fatto nella periferia.

Le economie di scala valgono sia per gli investimenti privati sia per quelli governativi nei servizi pubblici. Certi servizi, come scuole, strade, ospedali, sistemi fognari e di acqua potabile, forze di polizia e protezione antincendio, costituiscono quella che è nota col nome di infrastruttura dell'industria: si tratta cioè di servizi che non entrano direttamente nei processi di produzione, ma senza i quali l'industria non potrebbe esistere. Per rendersi conto del modo in cui le economie di scala influiscono sugli investimenti pubblici si consideri la situazione di un paese in via di sviluppo che debba decidere se costruire una nuova scuola superiore nel centro o nella periferia. Il costo dell'edificio scolastico sarebbe approssimativamente lo stesso in entrambe le zone. Nel centro però vi sono già autobus per portare gli alunni a scuola, strade su cui gli autobus possono correre, un buon numero di insegnanti qualificati e di scuole di grado inferiore per preparare gli alunni all'istruzione superiore.

Alla periferia invece bisognerebbe creare, se non tutti, almeno alcuni di questi elementi prima che la prima classe arrivi al diploma. Per questa ragione il suo costo sarebbe notevolmente più elevato alla periferia di quanto lo sarebbe al centro. Argomenti analoghi valgono per altri servizi pubblici. Le disparità geografiche in fatto di infrastrutture sono particolarmente significative nelle prime fasi dello sviluppo economico, quando i capitali accumulati in una società sono ridotti.

Anche le economie di scala per gli investimenti privati sono notevoli nel centro. Nel centro emergente esiste una numerosa manodopera qualificata e vi sono efficienti sistemi di trasporto sia per i lavoratori sia per le merci. Agli inizi del periodo industriale, quando predominava l'industria pesante, avevano particolare importanza le strade ferrate per il trasporto delle materie prime e dei prodotti lavorati: per questa ragione vi sono molte più linee

Questa cartina presenta il centro degli Stati Uniti e quello del Canada. Le regioni centrali generalmente accettate dai geografi e dagli economisti figurano in colore intenso. Il centro degli Stati Uniti comprende le più antiche zone industriali degli stati nordorientali e medio-occidentali. Il centro del Canada è una zona industriale che confina con gli Stati Uniti. L'autore dell'articolo si è basato sui dati ufficiali della migrazione interna di 22 paesi. I dati sono aggregati per unità amministrative, come stati, province e prefetture. Le unità amministrative ufficiali sono state combinate in modo da definire con una certa approssimazione il centro, qui indicato in colore chiaro. Negli Stati Uniti il centro, secondo questo criterio, comprende le regioni chiamate nel censimento North Central e Northeast; la periferia è costituita dalle regioni South e West del censimento. Nel Canada il centro è la regione Center; le province atlantiche e il West sono la periferia. Le approssimazioni includono così un'area più vasta del centro vero e proprio. Il fatto che vi sia una migrazione netta dalle zone più estese indica che il brusco mutamento intervenuto nella migrazione è un'inversione di tendenza e non solo un riversamento dal centro verso zone limitrofe.

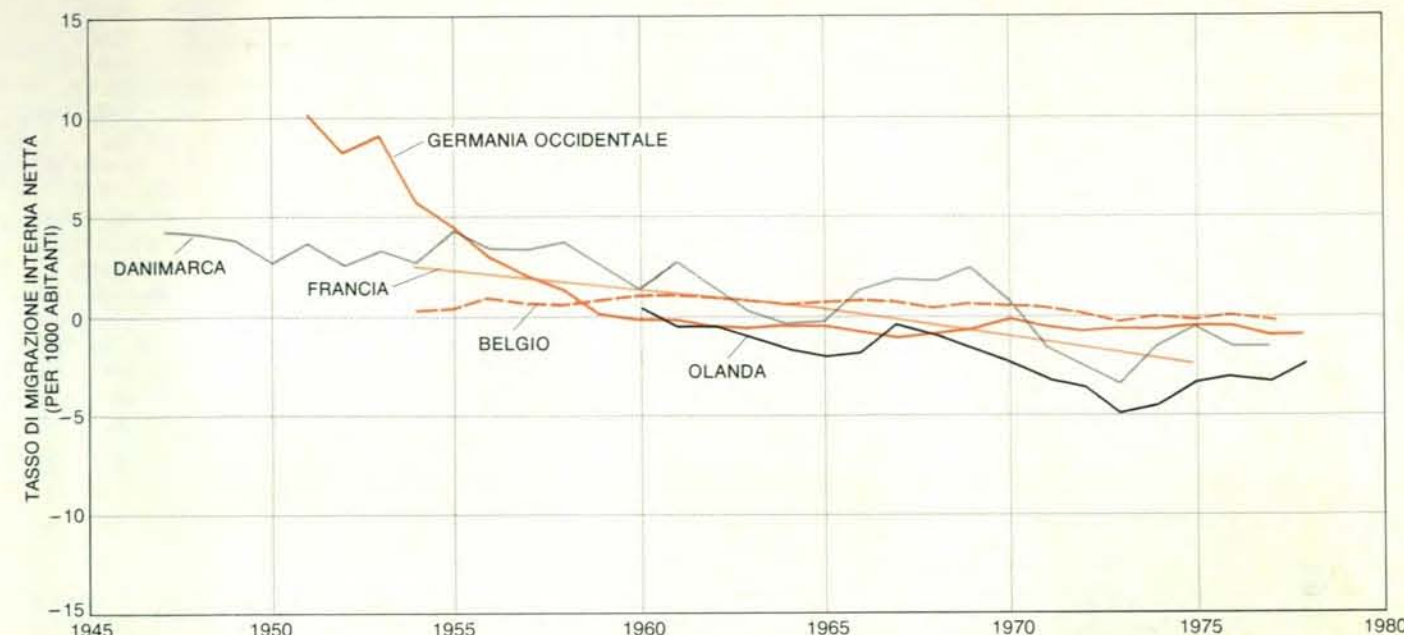
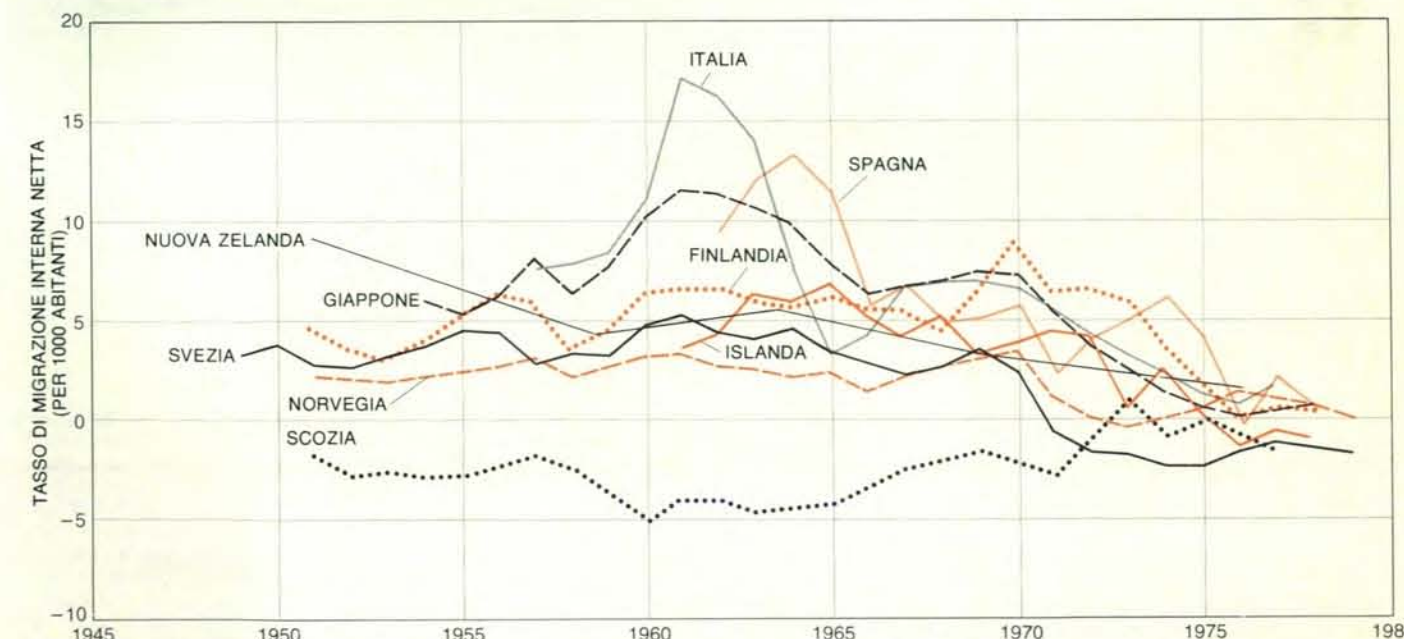
I tassi di migrazione relativi alle varie zone degli Stati Uniti vengono dati qui in termini di afflusso o efflusso per ogni 1000 abitanti di ogni regione del censimento. Dove il tasso netto è positivo la zona ha acquistato popolazione in virtù della migrazione, dove è negativo la zona ha perso abitanti. Gli Stati Uniti sono stati il primo paese in cui è mutata la direzione della migrazione. Dal 1945 il centro del paese ha perso costantemente popolazione a favore della periferia. Attualmente la perdita è di circa 400 000 persone l'anno. Poiché le informazioni del censimento del 1980 relative alla migrazione interna non sono ancora note, il grafico si basa su altri dati, che comprendono gli effetti dell'immigrazione dall'estero. Le perdite effettive del centro degli Stati Uniti a favore del South e del West sono probabilmente più consistenti. Agli inizi dello sviluppo economico l'industria si concentra nel centro per i vantaggi economici offerti dalla possibilità di costruire una fabbrica in una zona industriale già esistente. Tali vantaggi, che costituiscono le economie di scala, derivano in parte anche dalla vicinanza degli sbocchi commerciali e del mercato della manodopera e da efficienti sistemi di trasporto e di comunicazione. La popolazione segue l'industria. In una fase successiva dello sviluppo le economie di scala si riducono; l'industria e la popolazione incominciano allora a fluire verso la periferia. Una delle ragioni del grande e precoce flusso in uscita dal centro degli Stati Uniti è costituita dal fatto che South e West sono adatti all'industria.

I tassi di migrazione relativi alle regioni canadesi presentano un'inversione nell'ultimo decennio. A differenza dei dati sugli Stati Uniti, qui è inclusa solo la migrazione interna. Negli anni sessanta il centro canadese ha acquisito popolazione dalla periferia. Verso il 1970 però i centri industriali hanno incominciato a perdere abitanti a favore della periferia. Come negli Stati Uniti, la periferia è adatta allo sviluppo industriale. Quando i miglioramenti nel campo dei trasporti e delle comunicazioni e i cambiamenti nei processi industriali resero possibile l'installazione di impianti industriali e di altro genere fuori del centro, la periferia crebbe rapidamente. In Canada gran parte della migrazione interna è sempre stata diretta verso il West, che si va sviluppando sul piano economico. Di recente però anche le province atlantiche, per molto tempo «rimaste indietro» dal punto di vista economico e demografico, hanno incominciato ad attirare persone dal centro.

ferroviarie nel centro di quante ve ne siano alla periferia. Inoltre i centri urbani costituiscono un grande mercato tanto per i beni di consumo quanto per i beni strumentali e vi si trova con facilità tutta una gamma di imprese secondarie, fra cui non solo i fornitori di materie prime e di pezzi, ma anche studi legali, servizi di assistenza tecnica e ditte di pubbliche relazioni. Quando i sistemi di comunicazione erano limitati, la reciproca vicinanza delle grandi imprese nel centro offriva l'opportunità di un contatto costante fra i capi d'azienda.

Come conseguenza delle economie di scala, a mano a mano che una nazione si industrializza quasi tutti i nuovi impianti industriali vengono costruiti nel centro. La presenza dell'industria attira le aziende ausiliarie. La popolazione del centro incomincia ad aumentare via via che i lavoratori vengono attirati dagli alti salari offerti. Data la maggiore densità della popolazione, anche gli investimenti pubblici sono maggiori nel centro. A un certo livello di sviluppo, la crescita economica e demografica del centro diventa tale da autorafforzarsi.

La causa di fondo della distribuzione irregolare della popolazione in un paese sviluppato è quindi soprattutto di natura economica. L'americana Adna Weber, un'esperta di statistica che lavorava a New York come funzionaria statale, fu una delle prime persone a studiare la crescita delle aree urbane dal punto di vista statistico. Ecco la sua succinta formulazione del meccanismo della crescita demografica nelle zone industriali: «Quando l'organizzazione industriale richiede la presenza di manodopera in particolari località per accrescere la propria efficienza, la manodopera accorre sul posto, attirata da "una vita migliore...". Le forze economiche sono quindi la causa principale della concentrazione della popolazione nelle città». Che la concentrazione demografica e industriale dia luogo in una società all'utilizzazione più produttiva delle risorse e quindi a una maggiore ricchezza per la nazione è diventato un postulato fondamentale dell'economia urbana e della geografia economica. Il concetto è stato messo in rilievo da molti studiosi, fra cui Edward L. Ullman dell'Università di Washington, William Alonso della Harvard University, Colin Clark dell'Università del Queensland in Australia e Koichi Mera dell'Università di Tsukuba in Giappone.

Pare però che a un certo punto del processo di sviluppo economico i vantaggi delle economie di scala incomincino a diminuire. Sembra che questa diminuzione abbia avuto luogo prima di tutto negli Stati Uniti e anche nella maniera più accentuata. Come ho avuto modo di rilevare in precedenza, il centro degli Stati Uniti è costituito dagli stati industriali nordorientali e medio-occidentali, e questo è stato il primo centro di una nazione sviluppata ad avere una perdita netta di popolazione a beneficio della periferia come conseguenza della migrazione. È dal 1945 che negli Stati Uniti il centro ha un efflusso netto nel suo scambio con la periferia.

Fino al 1970, negli Stati Uniti, la perdita netta di popolazione dal centro alla periferia è stata in genere ridotta. Negli anni settanta però la migrazione è aumentata rapidamente fino a raggiungere un tasso di almeno 400 000 unità l'anno. Questo tasso, che è il più elevato dei paesi sviluppati, è stato calcolato sulla base delle stime della popolazione fatte dal Bureau of Census fra un censimento e l'altro. Purtroppo (ai fini della geografia economica) gli Stati Uniti non tengono un registro della popolazione. Sebbene nei censimenti decennali si raccolgano informazioni particolareggiate sulla migrazione

Cinque paesi industriali dell'Europa nordoccidentale presentano la flessione più precoce (all'infuori dell'America Settentrionale) della migrazione verso il centro. Ogni curva rappresenta il tasso di migrazione netta relativa al centro del paese. La prima flessione ha avuto inizio nella Germania Occidentale, a causa in parte della distribuzione relativamente regolare della popolazione in quel paese. Negli anni sessanta la migrazione era già diminuita in tutte e cinque le nazioni; negli ultimi anni settanta ne era derivato un efflusso netto dal centro. Le zone periferiche, come l'Olanda settentrionale, che da decenni continuavano a perdere popolazione a vantaggio del centro, stanno ora acquistando nuovi abitanti in virtù della migrazione. Una delle ragioni di tale efflusso è che la periferia di questi paesi ha aree adatte per l'industria.

Altri paesi dell'Europa occidentale, unitamente al Giappone e alla Nuova Zelanda, hanno un modello migratorio diverso da quello dei paesi industriali dell'Europa nordoccidentale. In molti dei paesi esterni dell'Europa il tasso di migrazione verso il centro raggiunse il culmine negli anni sessanta e incominciò a calare in misura significativa soltanto negli anni settanta. Inoltre, nella maggior parte delle nazioni di questo gruppo non si ha ancora nessun efflusso netto di persone dal centro. In alcuni paesi (come la Nuova Zelanda) le economie di scala sono ancora notevoli. In altri (come la Norvegia) la periferia non è molto adatta all'industria. In Gran Bretagna i dati completi sulla migrazione interna dopo la seconda guerra mondiale sono disponibili soltanto per la Scozia. Dato però che la Scozia costituisce una parte significativa della periferia, il tasso di migrazione da questo paese può essere usato come immagine speculare del movimento verso il centro, situato nel sud-est.

interna, non sono ancora disponibili i risultati relativi al 1980. Inoltre, fonti diverse dal censimento forniscono solo un resoconto approssimativo dei luoghi in cui gli immigranti dall'estero si insediano negli Stati Uniti. Per questa ragione le informazioni sulle tendenze demografiche degli anni settanta delle varie regioni degli Stati Uniti comprendono non solo gli effetti della migrazione fra regione e regione, ma anche gli effetti dell'immigrazione.

La cifra di 400 000 persone all'anno è quindi senza dubbio alcuno una sottovalutazione della migrazione netta effettiva dal centro americano. L'immigrazione nel nord-est dall'estero è notevole, e probabilmente è andata sempre più aumentando nell'ultimo decennio. Quando saranno disponibili i dati particolareggiati del censimento e sarà possibile separare gli effetti dell'immigrazione da quelli della migrazione interna, la perdita netta del centro risulterà probabilmente superiore alle 400 000 unità l'anno.

Sulla consistenza numerica della popolazione di una zona influisce, oltre la migrazione interna e l'immigrazione, anche il tasso di incremento naturale, che è pari alla differenza fra il numero delle nascite e quello delle morti in un dato periodo. I tre processi demografici possono incrementarsi o attenuarsi a vicenda. Come vedremo, in alcuni paesi la riduzione della popolazione del centro dovuta alla migrazione è stata messa in ombra dal tasso di incremento naturale, che al centro è sempre più elevato della media del paese nel suo insieme. Inoltre le cifre relative alla popolazione totale di un paese sono disponibili molto più rapidamente dei dati sulla migrazione interna; questa è l'unica ragione per la quale l'inversione dei flussi migratori ha attirato relativamente poca attenzione da parte degli studiosi di scienze sociali.

Il centro del Canada, così come l'ho definito nel mio lavoro, è costituito dalle province dell'Ontario e del Quebec. Queste province comprendono Ottawa, la capitale nazionale, nonché Toronto e Montreal, le città più importanti. Il centro effettivo del Canada, così come viene definito generalmente dai geografi e dagli economisti, è un'area più piccola che confina con gli Stati Uniti. Negli anni sessanta il centro canadese attirò popolazione dal resto del paese. Alla fine del decennio però il tasso di trasferimento nel centro subì una netta flessione, che nei primi anni settanta si trasformò in un flusso netto in direzione della periferia; questo efflusso continuò per tutti gli anni settanta.

Come negli Stati Uniti, così anche in Canada una parte consistente della migrazione interna è diretta verso la parte occidentale del paese. Oggi però anche le province della costa atlantica, che per lungo tempo hanno avuto un tasso ridotto di crescita demografica e grandi difficoltà economiche, stanno acquistando popolazione proveniente dal centro densamente popolato. Sebbene quasi tutti gli immigranti si insedino nell'Ontario e nel Quebec, la notevole migrazione netta in uscita dal centro ha dato luogo a un tasso di crescita demografica che è inferiore a quello della nazione nel suo insieme. In Canada pertanto la migrazione interna è diventata tanto rilevante da superare sia l'immigrazione sia l'incremento naturale.

Nei paesi molto industrializzati dell'Europa nordoccidentale la perdita del potere di attrazione del centro ha portato negli anni sessanta e settanta a una migrazione netta verso l'esterno. Il centro dell'Olanda, noto con il nome di Randstad, si trova sulla costa occidentale del paese. Da un po' di tempo a questa parte si va espandendo verso nord-est, in direzione cioè dei centri industriali della Germania Occidentale. Da parte di Hans Blumenfeld è stata avanzata l'ipotesi che la migrazione netta in uscita dal Randstad iniziata negli anni sessanta sia stata solo l'estensione verso sud-est del centro tradizionale. Dai miei dati risulta per altro che anche le zone più remote dell'Olanda stanno attualmente acquistando popolazione a detrimento del Randstad. Queste zone comprendono il nord e il sud-ovest, che per molti decenni sono «rimasti indietro» tanto sul piano economico quanto su quello demografico.

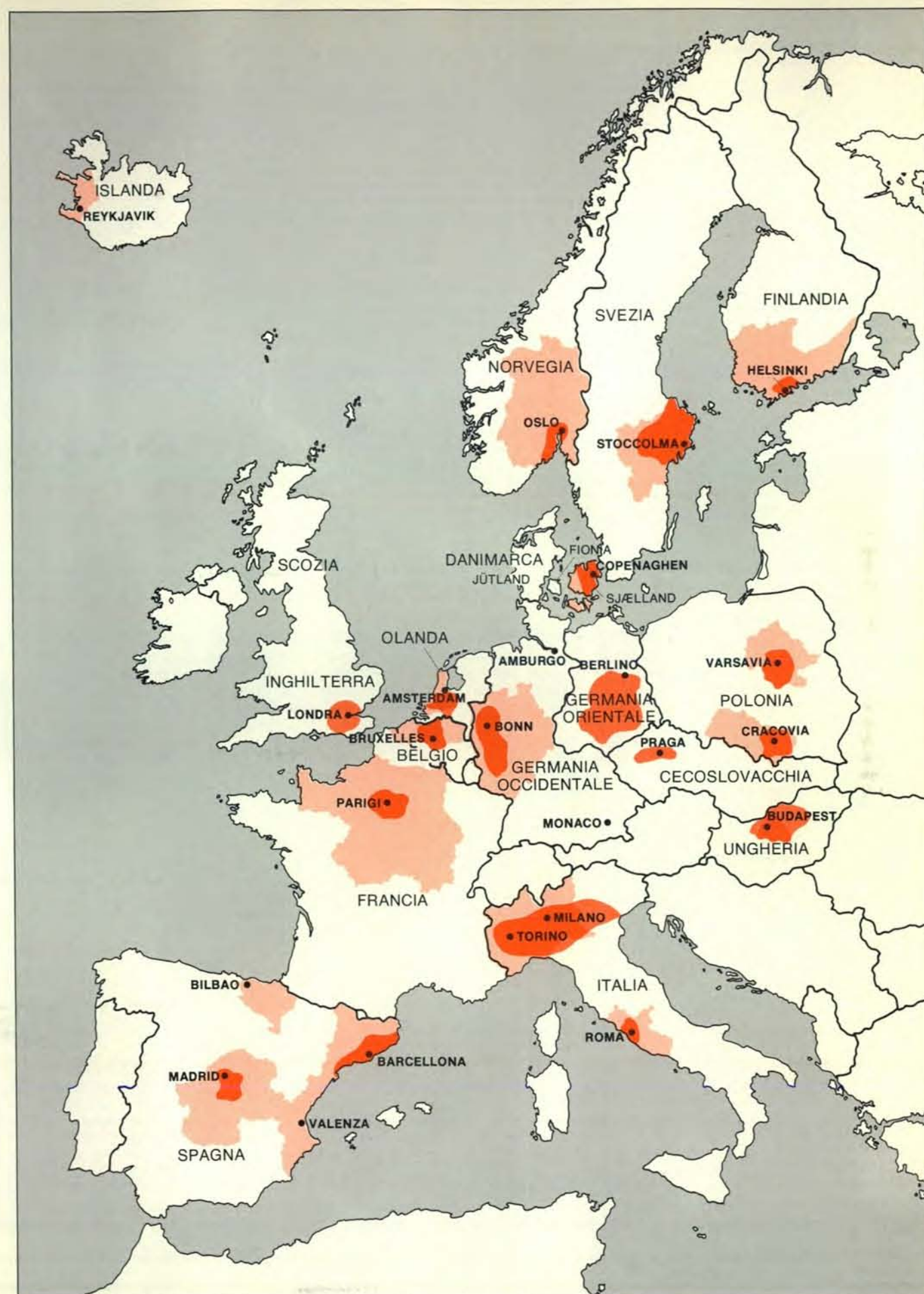
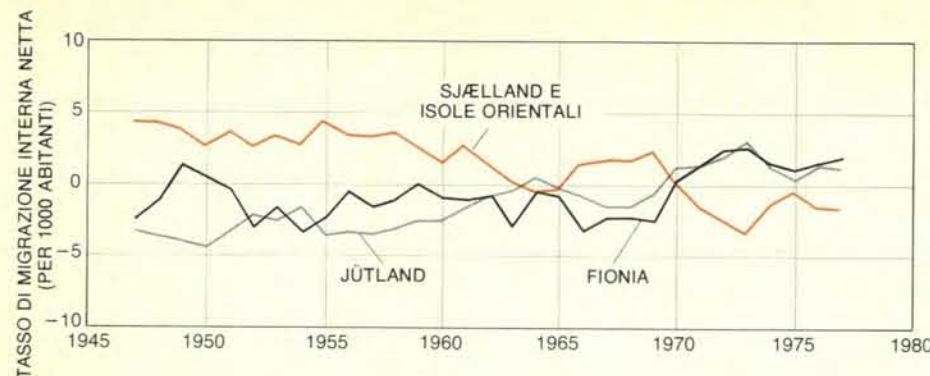
Analogamente, anche in Danimarca, in Belgio e in Francia si è avuta di recente un'inversione di tendenza dei precedenti flussi migratori. In Danimarca la migrazione interna è sempre stata diretta verso oriente, ossia dallo Jütland e dalla Fionia rurali verso Copenaghen e le zone limitrofe dell'isola Sjælland; ora invece c'è un flusso notevole nella direzione opposta. In Belgio la provincia settentrionale delle Fiandre, densamente popolata, sta perdendo ora residenti a favore della Vallonia, la provincia meridionale meno densamente popolata. Le Fiandre comprendono la zona urbana intorno ad Anversa, Bruxelles e Gand che è stata il centro demografico del Belgio.

In Francia il censimento del 1975 ha rivelato che dal 1968 al 1975, per la prima volta da quando aveva avuto inizio la prassi delle registrazioni, la regione e il bacino di Parigi hanno perso a favore del resto della Francia più popolazione di quanta ne abbiano acquisita. Le province mediterranee del sud della Francia hanno attirato popolazione dal 1945. Nei primi anni settanta però anche la parte occidentale della Francia, prevalentemente rurale, incominciò a registrare un flusso migratorio proveniente dalla parte orientale del paese, prevalentemente urbana, un fenomeno senza precedenti nella storia francese da prima dell'inizio dell'era industriale. Ciononostante, tanto in Francia quanto in Belgio il centro continua ad attirare una parte sproporzionata di immigranti dall'estero; la popolazione del centro ha inoltre un tasso di incremento naturale più elevato di quello della popolazione della periferia. In genere, pertanto, la frazione della popolazione del centro ha continuato ad aumentare. Le stime più recenti effettuate fra un censimento e l'altro fanno ritenere però che in Francia la popolazione del centro abbia incominciato a calare come percentuale della popolazione nazionale.

Nella Germania Occidentale il centro è costituito dalla zona industriale densamente popolata compresa tra i fiumi Reno e Ruhr. Questa regione cominciò a perdere popolazione a favore del resto della Germania Occidentale negli anni sessanta, prima che l'inversione fosse in atto negli altri paesi industriali d'Europa. Sebbene la densità demografica del distretto Reno-Ruhr sia molto elevata, la popolazione della Germania Occidentale è distribuita in modo un po' più regolare di quella della maggior parte degli altri paesi industriali; le due città più grandi del paese, Amburgo e Monaco, si trovano nella periferia. Non è da escludere che la distribuzione relativamente regolare spieghi in parte la precoce inversione della migrazione netta.

Come ho detto prima, la concentrazione dell'industria e della popolazione nel centro è dovuta in grande misura alle economie di scala. Il fatto che nelle nazioni sviluppate la migrazione verso il centro sia stata sostituita da un flusso verso la periferia fa ritenere che le economie di scala si siano ridotte. Parecchi fattori potrebbero aver concorso a ridurle. Forse quello più importante è stato lo sviluppo tecnologico, che ha mutato i criteri di valutazione per decidere dove si debbano costruire gli impianti industriali.

Da qualche decennio gli investimenti

Il centro dei paesi europei comprende nella maggior parte dei casi non solo le zone metropolitane più importanti, ma anche la capitale. Il centro vero e proprio è rappresentato in colore intenso, il centro approssimativo, composto di unità amministrative, in colore chiaro. Nel campione studiato dall'autore dell'articolo i paesi europei si possono dividere in tre gruppi sulla base dei loro modelli migratori; questi raggruppamenti sono in larga misura anche geografici. Al primo gruppo appartengono cinque paesi industriali dell'Europa nordoccidentale: Belgio, Danimarca, Francia, Germania Occidentale e Olanda. In questi paesi c'è attualmente un flusso netto di persone in uscita dal centro. La zona delimitata dai fiumi Reno e Ruhr, il centro della Germania Occidentale, è la preminente concentrazione industriale d'Europa. La popolazione della Germania Occidentale per altro è distribuita in maniera più regolare di quella di altre nazioni industriali: Amburgo e Monaco, le città più grandi, si trovano nella periferia. Il secondo gruppo comprende gli altri paesi dell'Europa occidentale (Finlandia, Inghilterra, Islanda, Italia, Norvegia, Spagna e Svezia) nonché il Giappone e la Nuova Zelanda. In queste nazioni il tasso di migrazione verso il centro è diminuito, ma soltanto fino a un livello tale da creare un certo equilibrio fra le migrazioni in entrata e in uscita dal centro. Reykjavik è il vero centro dell'Islanda. Il terzo gruppo comprende quattro paesi dell'Europa orientale: Cecoslovacchia, Germania Orientale, Polonia e Ungheria. Qui c'è un flusso verso il centro. Berlino Est è il centro della Germania Orientale e Praga quello della Cecoslovacchia, così come sono stati definiti approssimativamente dall'autore dell'articolo.

La migrazione fra una zona e l'altra della Danimarca manifesta negli anni settanta un'inversione della direzione prevalente della migrazione interna. Le più importanti aree territoriali della Danimarca sono l'isola orientale di Sjælland, sulla quale si trova Copenaghen, l'isola centrale di Fionia e la penisola occidentale dello Jütland. Per molti decenni le economie di scala furono notevoli e la popolazione si spostò prevalentemente verso est, ossia dallo Jütland rurale a Copenaghen. Secondo il criterio seguito dall'autore dell'articolo, la regione centrale comprende la Sjælland e parecchie isolette limitrofe, e quindi è più grande del centro vero e proprio. Dal 1970 in poi questa grande zona ha continuato costantemente a perdere popolazione a causa della migrazione. Nello stesso periodo lo Jütland ha incominciato ad attirare persone dal centro; anche l'isola di Fionia sta guadagnando abitanti. Oggi il flusso migratorio netto è diretto verso occidente e, per la prima volta in questo secolo, il tasso di crescita demografica nello Jütland è superiore a quello della Danimarca nel suo insieme. Analoghe inversioni sono state rilevate in Belgio, in Francia e in Olanda. Uno dei fattori che hanno contribuito a determinare queste inversioni è stato l'avvento di processi industriali basati su tecnologie avanzate. Molti prodotti di industrie avanzate sono facili da trasportare; il numero degli addetti è ridotto, ma gli impianti richiedono una superficie estesa. Risulta quindi vantaggioso costruire nuovi impianti in zone periferiche.

effettuati nei paesi sviluppati sono stati destinati in gran parte alla produzione di merci che incorporano tecnologia avanzata: ne sono un esempio i calcolatori elettronici di piccole dimensioni. I fattori che vengono presi in considerazione quando si tratta di decidere dove costruire questi calcolatori sono completamente diversi da quelli che all'inizio dell'industrializzazione inducevano a scegliere il luogo in cui costruire un'acciaieria. I componenti dei prodotti ad alta tecnologia sono generalmente piccoli e leggeri, così come lo sono i prodotti finiti. Con i miglioramenti intervenuti nei trasporti, e in particolare con la costruzione delle autostrade, è diventato pratico ed economico trasportare i componenti da luoghi lontani e portare i prodotti finiti su mercati altrettanto lontani.

Nella maggior parte di questi impianti il numero degli addetti è piuttosto ridotto e, di conseguenza, si riduce anche l'importanza di essere vicini a un mercato del lavoro ricco di manodopera qualificata. I miglioramenti intervenuti nelle comunicazioni hanno reso possibile che le industrie manifatturiere si insediassero lontano dalle ditte che forniscono servizi accessori. Una volta raggiunto un alto livello di sviluppo economico, l'accumulazione dei capitali è tale ormai da permettere la costruzione delle infrastrutture nelle regioni periferiche. Le scuole, le strade, gli ospedali e i corpi dei vigili del fuoco degli Stati Uniti sudoccidentali o della Francia occidentale non hanno niente da invidiare a quelli del centro. Per di più la crescita del settore dei servizi ha contribuito a ridurre i vantaggi economici del centro; i limiti imposti dalla geografia alla localizzazione delle società di servizi sono stati molto ridotti dai progressi compiuti nel campo delle comunicazioni.

La convenienza economica offerta dal centro rispetto alla periferia è stata pertanto ridotta da parecchi fattori. In un'area urbana molto densamente popolata, inoltre, alcune delle economie di scala si possono addirittura trasformare in diseconomie. Molti processi produttivi, per esempio, richiedono un edificio a un solo piano, tale da occupare una notevole area di pianura. Nelle aree urbane le superfici piane sono molto costose; nella periferia meno sviluppata i costi di costruzione sono notevolmente inferiori. Al centro le organizzazioni sindacali fanno salire le retribuzioni oltre il tasso prevalente nella periferia. L'elevata densità demografica può condurre alla congestione del traffico, che diventa un ostacolo per l'industria via via che il trasporto delle merci viene effettuato in misura sempre maggiore dagli autocarri. Il centro è caratterizzato poi anche da premi assicurativi elevati, da inquinamento ambientale e da incertezze sociali.

La riduzione delle economie di scala può tuttavia indurre l'industria e il commercio a spostare la propria ubicazione solo se la periferia dispone di aree adatte per costruirvi fabbriche, aziende e case d'abitazione. Gli Stati Uniti sono il miglior esempio di un paese con una periferia che è almeno pari, se non addirittura superiore, al centro per quel che riguarda il patrimonio naturale.

Nessuno degli altri paesi del mondo sviluppato ha una periferia con risorse non ancora utilizzate così ingenti come quelle degli Stati Uniti. Questa è la ragione principale per la quale nelle altre nazioni il tasso di migrazione dal centro è stato inferiore a quello riscontrato negli USA. La periferia dei paesi dell'Europa nordoccidentale ha, comunque, le aree adatte per l'industria e per l'insediamento umano. Il clima, il suolo e la topografia della Francia sudoccidentale, dello Jütland, dell'Olanda settentrionale, della Baviera e del Belgio meridionale sono adatti per lo sviluppo urbano e industriale; lo stesso si può dire per le province occidentali del Canada. Finché le economie di scala si mantenevano consistenti, queste zone ristagnavano dal punto di vista economico. Da quando i vantaggi di scala si sono ridotti o si sono addirittura trasformati in svantaggi, le regioni periferiche hanno incominciato a esercitare una forte attrattiva sia sulle imprese commerciali ad alto rischio sia sulla popolazione.

I nove paesi che compongono il terzo gruppo del mio campione (Finlandia, Giappone, Inghilterra, Islanda, Italia, Norvegia, Nuova Zelanda, Spagna e Svezia) presentano un modello diverso da quello dell'America Settentrionale e dell'Europa nordoccidentale. In queste nazioni il tasso migratorio verso il centro toccò la sua punta massima negli anni sessanta, un po' più tardi che negli altri gruppi, e incominciò a diminuire in misura significativa soltanto negli anni settanta. Una distinzione più importante è costituita dal fatto che in Giappone, nella Nuova Zelanda e nei paesi esterni dell'Europa occidentale la flessione del tasso di migrazione non ha portato, nella maggior parte dei casi, a una perdita netta della popolazione del centro. Il flusso migratorio ha mostrato invece la tendenza a diminuire per poi stabilizzarsi di colpo, lasciando o un equilibrio fra il movimento in entrata e quello in uscita dal centro o una piccola migrazione netta in entrata.

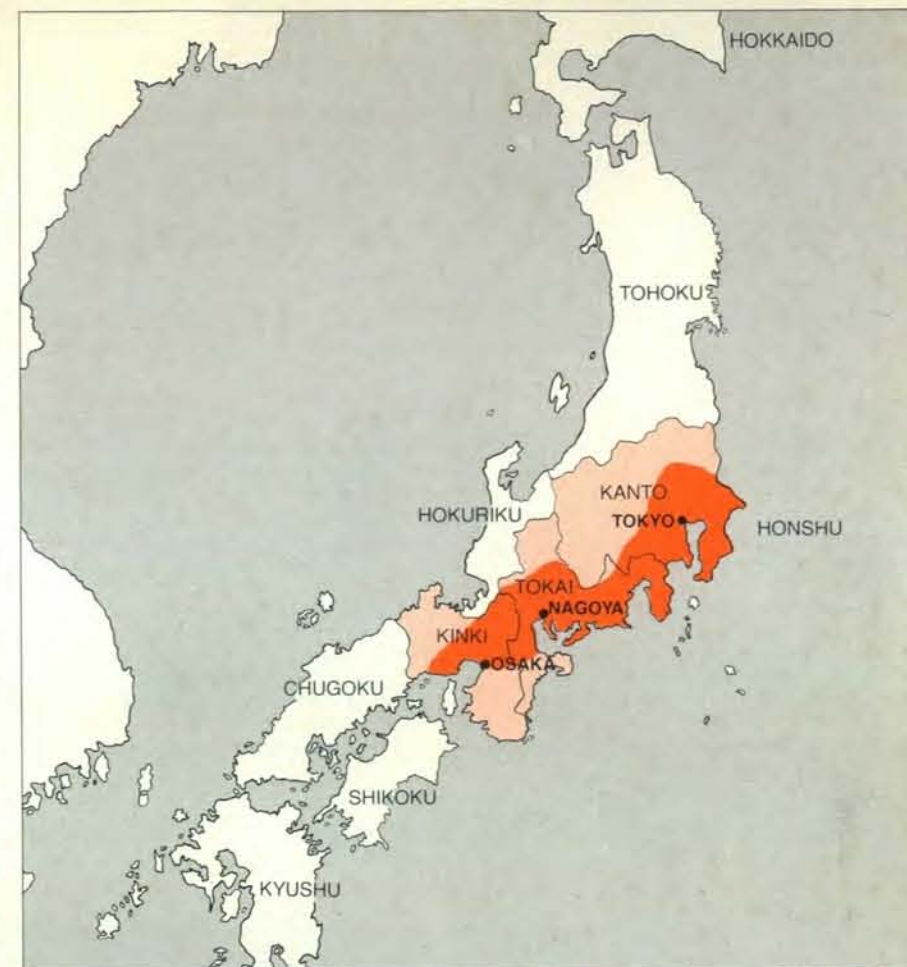
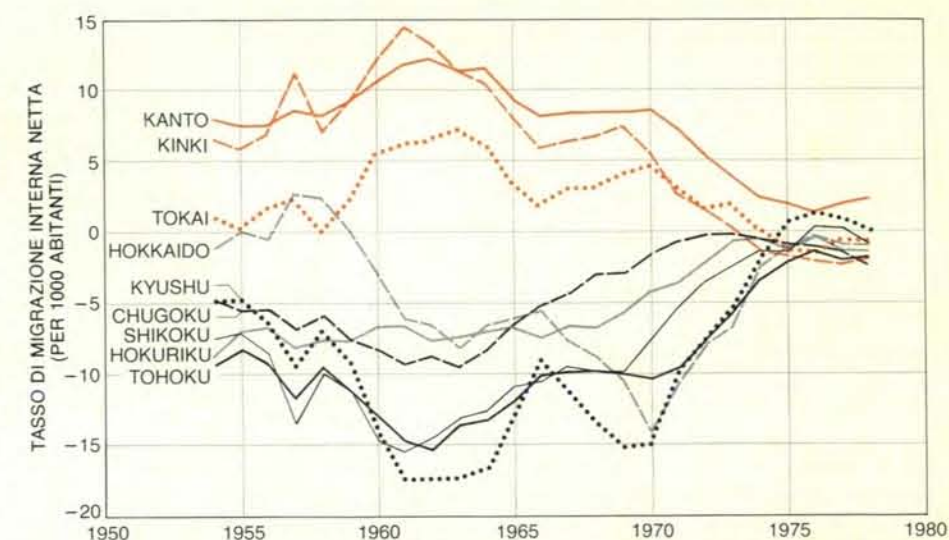
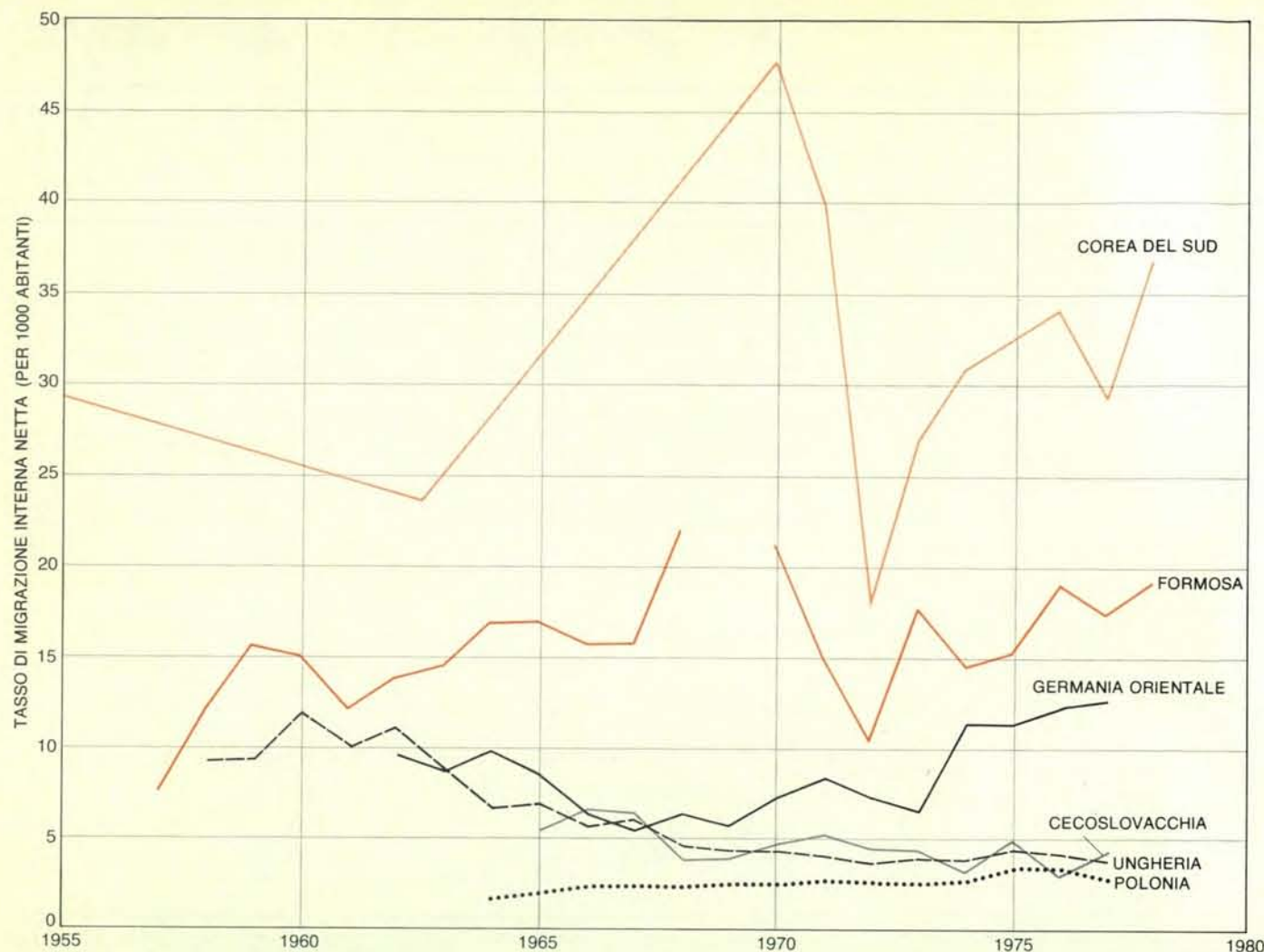
Nei paesi di questo terzo gruppo il tasso complessivo di crescita demografica del centro rimane superiore a quello della nazione nel suo insieme. Ciononostante, non va sottovalutata l'importanza della diminuzione della migrazione netta in paesi come l'Italia, il Giappone e la Svezia. In queste nazioni i geografi, gli economisti e i pianificatori urbani avevano pensato che sarebbe stato impossibile opporsi al flusso della popolazione verso il centro densamente popolato e interrompere lo spopolamento della periferia. I dati da cui risultava che ci si stava avvicinando o che addirittura si era arrivati a un certo equilibrio furono accolti con stupore e in alcuni casi con scetticismo.

Perché nei paesi del terzo gruppo la direzione della migrazione netta non si è invertita, così come è avvenuto nell'America Settentrionale e nell'Europa nordoccidentale? Con ogni probabilità la ragione va ricercata nel fatto che, al di fuori delle zone industriali, vi è scarsità di aree adatte allo sviluppo. In ognuno di questi paesi (con la possibile eccezione della Nuova Zelanda) la periferia è in evidente svantaggio geografico rispetto al centro. Nei paesi scandinavi la periferia non ha pianure estese e quelle che ha sono frammentate in piccoli appezzamenti da montagne e da fiordi. Nelle regioni settentrionali l'estate è fredda e umida e l'inverno è lungo, buio e freddissimo. Queste condizioni rendono difficile costruire città, fabbriche e strade e attirare e mantenere in luogo forze di lavoro adeguate.

Il Mezzogiorno d'Italia, che costituisce buona parte della periferia della nazione, ha una topografia difficile e disponibilità idriche incerte. In Giappone gli inconvenienti delle regioni periferiche sono estremi. C'è una particolare mancanza di terreni pianeggianti al di fuori della parte centrale dell'isola di Honshu, su cui sorgono tutte le città giapponesi più importanti. Gli unici grandi tratti di terreno pianeggiante disponibile sono nell'Hokkaido, l'isola più settentrionale, che ha un clima rigido con inverni freddissimi.

Come ho fatto notare prima, i vantaggi economici che derivano dall'insediamento di un'industria o di una società di servizi nel centro di un paese sono stati in gran parte annullati dalla crescita economica e demografica. Esistono però ancora alcuni svantaggi nell'essere lontani dalle principali concentrazioni di industrie e di popolazione. Quando questi svantaggi si uniscono a un terreno o a un clima difficili, un rapido sviluppo della regione periferica è chiaramente da escludere. L'unica eccezione a questo modello, per quel che riguarda il terzo gruppo di nazioni, è la Nuova Zelanda, dove il centro è costituito dalla parte settentrionale dell'Isola del Nord e la periferia comprende il resto dell'Isola del Nord e l'Isola del Sud. Quest'ultima non ha gravi inconvenienti di natura geografica. La più importante concentrazione industriale intorno a Auckland nell'Isola del Nord, però, è di dimensioni piuttosto ridotte e, quindi, le economie di scala all'interno del paese rimangono consistenti.

Dal momento che le regioni periferiche dell'Europa mediterranea, del Giappone e della Scandinavia hanno caratteristiche naturali così sfavorevoli, ci si chiede come mai il tasso di migrazione verso il centro abbia subito negli anni settanta una flessione tanto marcata. Una ragione va ricercata nel fatto che il venir meno delle economie di scala è stato accompagnato da politiche governative volte a impedire un ulteriore efflusso di popolazione dalla periferia. In Norvegia, per esempio, il governo ha attuato un programma che aveva l'esplicito intento di rendere le zone periferiche settentrionali più allettanti per gli individui e per le aziende. Le infrastrutture sono state migliorate a spese dello stato. Sono state costruite strade con riporto duro adatte a tutte le condizioni atmosferiche, sono stati costruiti ospedali e scuole ed è stato istituito un servizio di battelli molto veloci per collegare i villaggi costieri isolati. Inoltre sono state concesse sovvenzioni alle aziende che investono capitali nel nord.

Politiche di «deconcentrazione» di questo genere sono state attuate in molti dei 22 paesi del mio campione. Nell'America Settentrionale e nell'Europa nor-

La regione centrale del Giappone è presentata in colore intenso. Per definire approssimativamente il centro giapponese, l'autore ha unito fra loro le regioni di Kanto, Kinki e Tokai, indicate in colore chiaro. Queste tre zone si trovano nell'isola di Honshu e comprendono le principali città nipponiche: Tokyo, Nagoya e Osaka. Il centro del Giappone ha una popolazione di circa 65 milioni di abitanti. Una delle ragioni dell'elevata densità demografica al centro è costituita dal fatto che in Giappone quasi tutto il terreno pianeggiante adatto per le costruzioni si trova nella parte centrale di Honshu, sulla costa orientale. Nella periferia le uniche aree di pianura di una certa consistenza si trovano nell'isola settentrionale di Hokkaido, che ha un clima estremamente rigido.

In Giappone la migrazione fra una zona e l'altra è chiaramente limitata da fattori geografici. Le zone centrali di Kanto, Kinki e Tokai sono rappresentate dalle tre curve in alto. Le sei curve in basso si riferiscono alle zone della periferia. Come nei paesi esterni dell'Europa occidentale, la migrazione verso il centro toccò la punta massima negli anni sessanta e poi subì una brusca flessione negli anni settanta. In Giappone però non c'è stato un efflusso netto dal centro dovuto alla migrazione. La ragione principale è che condizioni poco favorevoli di natura geografica rendono le regioni periferiche inadatte allo sviluppo industriale e all'insediamento urbano su grande scala.

L'Europa orientale e il Terzo Mondo presentano modelli migratori affatto diversi da quelli dei paesi più sviluppati. In Cecoslovacchia, nella Germania Orientale, in Ungheria e in Polonia c'è un flusso piuttosto ridotto, ma costante di persone verso il centro, nonostante le politiche governative volte a scoraggiare un'ulteriore migrazione dalla periferia. Nella Corea del Sud e a Formosa lo sviluppo economico è ancora in corso e il tasso di migrazione verso il centro è estremamente elevato. Ogni anno la zona di Seul guadagna popolazione in misura pari al 3-4 per cento dei suoi abitanti; l'area di Taipei ne guadagna quasi il 2 per cento. Questi tassi, fra i più elevati mai registrati, riflettono il fatto che nel Terzo Mondo le economie di scala sono ancora notevoli. Dato un mutamento intervenuto nel modo di elaborare le statistiche, i dati relativi alla migrazione interna fra una zona e l'altra di Formosa nel 1969 non sono omogenei con i dati presentati, e quindi sono stati omessi.

doccidentale esse sono servite semplicemente ad accrescere l'attrattiva delle regioni periferiche. Nei paesi del terzo gruppo, d'altra parte, le politiche di deconcentrazione hanno contribuito a compensare alcune carenze delle regioni periferiche. L'equilibrio dei flussi migratori in questi paesi rimane delicato. In alcuni di essi (fra cui l'Inghilterra, l'Italia e il Giappone) si è avuto di recente un aumento del tasso di migrazione in direzione del centro, anche se tali aumenti non hanno portato il tasso in questione al livello degli anni cinquanta e sessanta. Dubito che nel mondo sviluppato, fatta eccezione per l'America Settentrionale e per l'Europa continentale nordoccidentale, possa mai aver luogo un movimento consistente in direzione della periferia. Negli altri paesi sembra probabile che si arrivi al massimo a un equilibrio tra i flussi.

Forti politiche di deconcentrazione sono state adottate anche nell'Europa orientale. Queste politiche però non sono state così efficaci come nell'Europa occidentale, perché i paesi orientali sono a un livello più basso di sviluppo economico. Le economie di scala sono ancora notevoli e per di più l'accumulazione di un sovrappiù di capitali è ancora troppo ridotta per permettere al governo di offrire grandi sovvenzioni alla periferia. Di conseguenza in Cecoslovacchia, nella Germania Orientale, in Polonia e in Ungheria c'è un tasso netto abbastanza consistente di migrazione verso i grandi insediamenti urbani. Ciononostante, questo tasso è inferiore a quelli registrati nell'Europa occidentale quando il centro si andava rapidamente espandendo.

Nel mondo in via di sviluppo le economie di scala sono un incentivo alla concentrazione ancor più di quanto lo siano nell'Europa orientale; esse corrispondono alle economie che prevalevano nel mondo sviluppato in una fase d'industrializzazione molto anteriore. Inoltre, pur essendo state adottate in alcune parti del mondo in via di sviluppo, le politiche di deconcentrazione non hanno la benché minima forza. Di conseguenza la migrazione netta verso il centro è molto elevata. La Corea del Sud e Formosa sono le uniche parti del Terzo Mondo per le quali siano disponibili statistiche annuali sulla migrazione interna. In entrambi i luoghi il flusso verso il centro è enorme e non dà segni di cedimento. Il tasso di migrazione nell'area di Seul, il centro della Corea del Sud, è compreso fra il 3 e il 4 per cento all'anno; per Taipei, il centro di Formosa, il tasso varia dall'1 al 2 per cento. Questi tassi sono notevolmente superiori a quelli dell'Europa o degli Stati Uniti in una fase comparabile di sviluppo; si tratta in realtà di tassi di migrazione netta in direzione del centro fra i più elevati che siano mai stati registrati. I dati di recenti censimenti effettuati in altri paesi in via di sviluppo, fra cui Brasile, India, Messico e Turchia, fanno ritenere che anche lì la migrazione verso le grandi città non sia diminuita.

Pare che sia necessario un alto livello di sviluppo economico prima che la deconcentrazione demografica possa incominciare. Va notato che il livello di sviluppo economico e il tasso di migrazione dal centro non sono in rapporto diretto con la popolazione delle aree urbane più grandi. Il flusso di persone verso il centro dell'Islanda, della Norvegia, dell'Italia e del Giappone registrò intorno al 1970 una repentina flessione e in ogni singolo paese il tasso di migrazione netta si sta attualmente avvicinando allo zero. Si consideri la disparità in fatto di consistenza numerica fra i centri di questi paesi: l'Islanda sudoccidentale ha una popolazione di 150 000 abitanti, la Norvegia orientale di due milioni, l'Italia nordoccidentale di 15 milioni e la zona di Tokaido in Giappone di 65 milioni.

Sebbene l'inizio della flessione della migrazione netta abbia un rapporto maggiore con il livello di sviluppo economico di quanto non ne abbia con la popolazione del centro, sembra proprio che il centro debba avere necessariamente una certa consistenza assoluta perché la flessione si trasformi in una significativa perdita di popolazione. Inoltre, come abbiamo visto, occorre che la periferia offra aree adatte per l'industria. Basta che manchi una sola di queste condizioni perché non si abbia nessun afflusso netto significativo. In Giappone, dove il centro ha una popolazione numerosa, ma la periferia presenta varie carenze dal punto di vista geografico, non vi è nessun flusso netto verso l'esterno; nella Nuova Zelanda, dove la periferia è vantaggiosa sul piano geografico, ma il centro è piccolo, non vi è nessun flusso verso l'esterno. In Francia, dove entrambe le condizioni sono soddisfatte, esiste un flusso migratorio dal bacino di Parigi verso la periferia.

L'inversione della migrazione interna nei paesi sviluppati è un fenomeno che non è ancora stato studiato a sufficienza. Qualunque ipotesi sulle sue cause è destinata quindi per forza di cose ad avere un carattere un po' congetturale. Anche se non è nota ancora in tutti i suoi aspetti, l'inversione potrebbe avere un grande effetto sul futuro politico del mondo sviluppato. Negli Stati Uniti, per esempio, ha già provocato uno spostamento nell'equilibrio del potere politico. I seggi alla Camera dei rappresentanti vengono distribuiti sulla base dei risultati del censimento. Dopo il censimento del 1980, per la prima volta in questo secolo il Sud e l'Ovest hanno alla Camera più seggi degli stati del centro. Eventuali spostamenti del potere politico in altri paesi sviluppati saranno probabilmente meno spettacolari di quello degli Stati Uniti. Ciononostante, l'attuale inversione dei modelli demografici porta indubbiamente con sé importanti conseguenze di natura sociale e politica che diventeranno sempre più evidenti nel resto di questo decennio e nei decenni che seguiranno.

Il moto browniano da Einstein a oggi

La teoria dinamica del moto browniano ci offre un filo conduttore entro la scienza fisica, dalla termodinamica statistica del non equilibrio alla teoria quantistica

di Bernard H. Lavenda

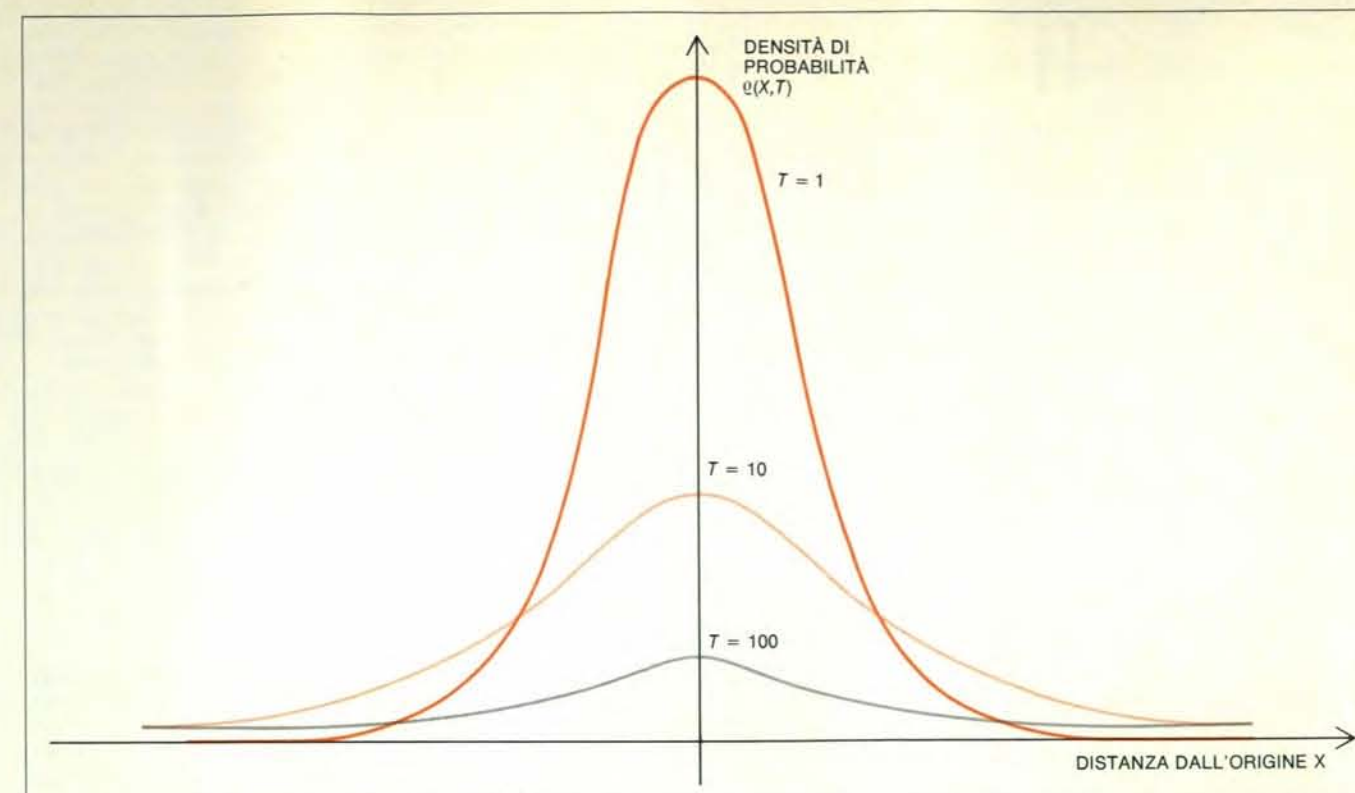
Nel 1827, il botanico inglese Robert Brown fu il primo a osservare al microscopio i minuscoli e rapidi moti irregolari di piccoli granelli di polline sospesi nell'acqua, cioè quello che oggi è conosciuto come «moto browniano». Il suo scopritore fu affascinato dal «rapido moto oscillatorio» dei granelli di polline e dal fatto che «l'inaspettata apparente vitalità di queste "molecole" persiste lungamente dopo la morte della pianta». Le prime spiegazioni attribuivano la causa dei moti irregolari alle correnti termiche convettive all'interno del fluido di sospensione: tuttavia, se così fosse, ci si aspetterebbe che il comportamento di una particella sia correlato con quello delle particelle vicine. Le osservazioni però non confermavano quest'idea, anzi il comportamento di una particella appariva indipendente dal suo passato. Questo dilemma fu indubbiamente causa di grande confusione, in un periodo in cui i pilastri della scienza erano fondati sui principi della meccanica classica: l'osservazione che il moto futuro è indipendente da quello passato, e che il moto è incessante, restò inspiegata per quasi un secolo.

Verso la fine dell'Ottocento si era già messo in evidenza che il moto browniano era tanto più rapido quanto più piccole erano le particelle e quanto più bassa la viscosità del fluido. Anche gli aumenti di temperatura provocavano un aumento della frequenza delle oscillazioni, così che in qualche modo la causa del moto doveva essere imputata ai moti termici delle molecole del mezzo. A quei tempi, la teoria cinetica dei gas era già stata sviluppata grazie al lavoro monumentale di James Clerk Maxwell e Ludwig Boltzmann, durante l'ultima metà del diciannovesimo secolo. Da tale teoria era noto che la temperatura di una sostanza è proporzionale all'energia cinetica media di agitazione delle molecole costituenti il mezzo. Se tale moto di agitazione potesse essere in qualche modo trasferito a molecole sufficientemente grandi da essere osservabili con un microscopio, ciò costituirebbe la prima evidenza diretta della validità della teoria cinetica del calore.

Nel periodo in cui apparve sulla scena Albert Einstein, nel 1905 (lo stesso anno in cui formulò la teoria della relatività ristretta, e in cui spiegò l'effetto fotoelettrico introducendo il concetto di «fotone»), era disponibile una notevole mole di dati sperimentali sul moto browniano. Ciononostante, Einstein era alla ricerca di fatti che comprovassero l'esistenza di atomi di dimensioni definite e non conosceva quel fenomeno; i suoi sforzi lo portarono a prevederne l'esistenza su basi puramente teoriche, e a fornirne la prima teoria quantitativa.

Consideriamo il moto di una particella libera, sulla quale, cioè, non agisce alcun campo di forze esterne: per determinare il moto della particella non basterebbe conoscere gli impulsi che essa riceve in un dato intervallo di tempo, ma occorrerebbe sapere anche la sua velocità iniziale. Per Einstein la conoscenza della velocità iniziale della particella browniana, rispetto a un qualsiasi intervallo di tempo di osservazione, rappresentava un dato insignificante in confronto al numero di urti che la particella riceve nello stesso intervallo di tempo. (In effetti la particella browniana subisce circa 10^{21} collisioni al secondo; quindi l'assunzione di Einstein è ampiamente giustificata.)

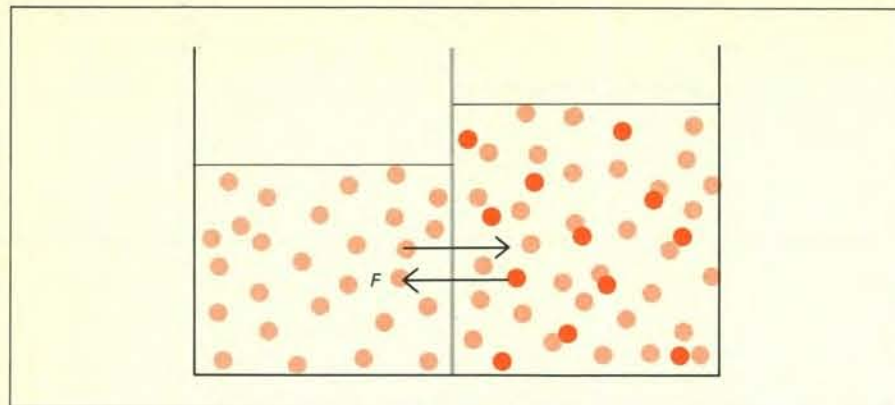
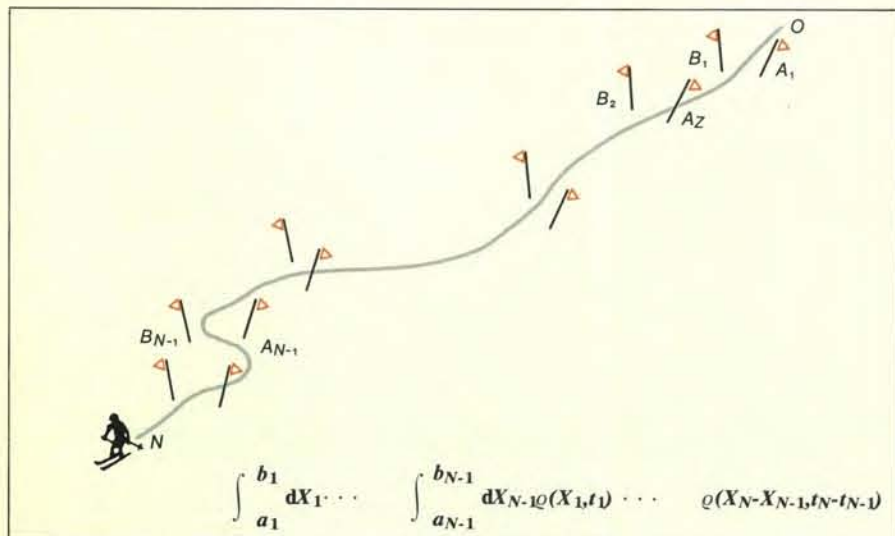
Facendo l'ulteriore assunzione di una distribuzione casuale delle posizioni molecolari e prendendo in considerazione intervalli di tempo lunghi rispetto all'intervallo di tempo medio fra due collisioni molecolari successive, Einstein giunse alla formulazione di un'equazione di diffusione, analoga a quella che descrive la conduzione termica, la cui soluzione fornisce la densità di probabilità che una particella browniana occupi una data posizione a un dato istante. Se consideriamo tale funzione nel caso di una particella che si trovi inizialmente nell'origine e diffonda in una sola direzione, la densità di probabilità si comporta in modo simile a una goccia di inchiostro che diffonde nel tempo in un bicchiere d'acqua. In altre parole, potremmo immaginare la densità di probabilità come la distribuzione, lungo una direzione dello spazio, di una sostanza estranea introdotta in un mezzo omogeneo, e l'evoluzione di questa densità di probabilità nel tempo come la diffusione della stessa sostanza. Per avvicinarci di più al modello reale, la posizione di un punto rappresentativo della sostanza a ogni dato istante dovrebbe essere descritta da una funzione casuale, associata al moto di ciascuna molecola della sostanza. La traiettoria deve avere necessariamente una forma estremamente complicata e discontinua, il che determina una caratteristica di vitale importanza del moto browniano: la velocità istantanea di un punto che descrive la traiettoria del processo non è definibile!

Comunque, in un intervallo di tempo finito, si può ottenere uno spostamento finito per il fatto che la velocità del punto rappresentativo inverte il suo segno con frequenza infinita, mentre il punto si muove in entrambe le direzioni. Così, dall'osservazione di un grande numero di processi casuali, otteniamo un gran numero di punti rappresentativi che si spostano in maniera erratica, o casuale, come in un movimento a «zig-zag», a causa delle interazioni con le particelle del mezzo. La pendenza di ogni tratto di cammino libero (lo «zig») non è necessariamente uguale a quella di un altro (lo «zag») e, con l'aumentare della frequenza delle interazioni, il «cammino libero medio» della particella diminuisce. Al limite dell'idealizzazione matematica, in cui il cammino libero medio tende a zero, possiamo dire che il vettore spostamento di una particella browniana non è differenziabile in alcun punto; non possiamo cioè definire una velocità per il processo. Conseguentemente, tutto quello che possiamo fare è parlare di una densità di probabilità, che equivale ad avere una densità di punti costituenti una sorta di gas che diffonde. Naturalmente queste particelle

La densità normale o gaussiana rappresenta la quantità $\varrho\,(X,t) = [4\pi Dt]^{-1/2} \times \exp(-X^2/4Dt)$. Possiamo immaginare un gran numero di particelle simili, affollate nelle immediate vicinanze di $X = 0$ all'istante $t = 0$, che vengono lasciate a se stesse da $t = 0$ in poi. Dopo un tempo t, si stabilisce spontaneamente una distribuzione di particelle tale che il relativo numero di particelle comprese fra X e $X + dX$ è : $\varrho(X,t)\,dX$. Oppure, alternativamente, possiamo considerare come nostro sistema, non un gran numero di particelle simili fra loro ma, piuttosto, una particella singola. Allora $\varrho(X,t)\,dX$ denota la probabilità che la particella si sia spostata, nel tempo t, in una regione compresa fra X e $X + dX$. Sulla base di questa formula, Einstein calcolò che l'allontanamento medio da $X = 0$ doveva essere $\sqrt{2Dt}$ al tempo t, in cui D è il coefficiente di diffusione. Einstein concluse così che il cammino descritto in media da una particella non è proporzionale al tempo, ma alla radice quadrata del tempo. Ciò deriva dal fatto che gli spostamenti descritti, per esempio durante due intervalli di tempo unitari, «non vanno sempre sommati fra loro, ma altrettanto frequentemente vanno sottratti».

di gas non possono essere né create né distrutte durante il loro moto e ciò equivale a dire che la probabilità si conserva; il valore della probabilità in una certa regione corrisponde al numero di punti che si trovano in quella regione oppure, equivalentemente, corrisponde al tempo che in media ogni «punto browniano» trascorre in quella regione.

Anche se questa descrizione costituisce un'astrazione limite dei processi che avvengono realmente in natura, è la sola che rende possibile allo stesso tempo l'introduzione di concetti probabilistici e l'interruzione della trasmissione di informazione tra passato e futuro. In altre parole, per descrivere il moto browniano, occorre simulare un processo senza «memoria», utilizzando quella che i matematici chiamano proprietà markoviana. Del resto, qualsiasi processo casuale, anche con memoria, può sempre essere decomposto in processi più elementari che godano della proprietà markoviana (ne troverete una prova vivente nella maggior parte delle istituzioni burocratiche esistenti).

Nella teoria di Einstein compare un solo parametro caratteristico - il coefficiente di diffusione - e Einstein derivò una formula in cui esso era espresso in termini del numero di Avogadro e di altre grandezze fisiche che potevano essere misurate in laboratorio. Einstein suppose che la diffusione delle particelle sospese nel liquido fosse governata da una condizione di «equilibrio dinamico» fra la forza osmotica che tende a sospingere le particelle dalle regioni a alta concentrazione alle regioni a bassa concentrazione, e una forza viscosa che tende a ritardare il moto delle particelle. La forza viscosa è proporzionale alla velocità della particella anziché alla sua accelerazione, poiché l'accelerazione iniziale subisce un rapido smorzamento in un mezzo viscoso. Einstein voleva evitare il ricorso alla nozione di velocità ben definita della particella e la novità del suo trattamento consiste nel tentativo di descrivere il moto delle particelle con ragionamenti probabilistici. Infatti l'analisi di Einstein prelude allo sviluppo della teoria matematica dei processi stocastici e l'eleganza del suo procedimento consiste nel fatto che la velocità introdotta con la forza viscosa è puramente virtuale! La formula di Einstein per il coefficiente di diffusione si applica anche quando non è definita alcuna velocità e anche quando vi è una sola particella browniana, per cui non è possibile definire la concentrazione! Allora, semplicemente invertendo la formula di Einstein, era possibile ottenere il numero di Avogadro dalla misura del coefficiente di diffusione di una sospensione colloidale di particelle sferiche di raggio approssimativamente uniforme. Considerando il numero di ipotesi introdotte nella derivazione della formula di Einstein, è veramente notevole che J. Perrin ottenesse un risultato sperimentale in accordo entro il 19 per cento col valore effettivo del numero di Avogadro, dedotto per altre vie.

Tutto ciò che abbiamo detto riguardo alla teoria di Einstein non le rende completa giustizia. Grazie agli sforzi di Einstein venne alla luce che la meccanica statistica era una teoria con implicazioni sperimentali che non potevano essere spiegate dalla termodinamica classica. La teoria delle fluttuazioni inaugurata dalla teoria di Einstein ha dato frutti che solo di recente cominciano a essere presi in considerazione dalle scienze fisiche e matematiche. Durante gli ultimi due decenni è stata sviluppata un'intensa ricerca, che nella scienza matematica va sotto il nome di «studio dei processi stocastici di diffusione». Sono state trovate immediate applicazioni nelle teorie di ottimizzazione dei controlli e del filtraggio dei segnali. Ma applicazioni ancora più ampie alle scienze fisiche e chimiche hanno portato a dimostrare che la teoria del moto browniano può sia costituire le fondamenta

Un esperimento sull'osmosi illustra la situazione di equilibrio «dinamico». Una membrana semipermeabile, permeabile alle molecole di solvente *(in colore chiaro)* e impermeabile alle molecole di soluto *(in colore più intenso)*, separa due compartimenti: uno contiene molecole di solvente soltanto e l'altro contiene sia molecole di solvente che di soluto. Il gradiente di concentrazione delle molecole di soluto dà luogo a una forza osmotica che agisce nel senso di aumentare il flusso di molecole di solvente nella direzione da sinistra verso destra. Nello stato di equilibrio dinamico, la forza osmotica F_O è bilanciata da una forza uguale ed opposta F_V, che è la forza viscosa che agisce in modo da ritardare il moto delle molecole di solvente. La differenza di livello fra i due compartimenti è uguale alla pressione osmotica del solvente.

Lo slalom (ottenuto con una simulazione al calcolatore) illustra il modo di Wiener di trattare il moto browniano come «somma su tutti i cammini». Sapendo che lo sciatore parte da O e arriva a N, la probabilità che esso passi attraverso ciascuna delle porte, negli istanti di tempo associati, è data dall'integrale multiplo riportato in basso a destra nell'illustrazione, in cui ϱ è la densità di probabilità visualizzata nell'illustrazione della pagina precedente (cioè la distribuzione trovata da Einstein). È piuttosto intuitivo che, se aumentassimo il numero di ostacoli e se avvicinassimo sempre più i paletti delle porte fra loro, potremmo tracciare il cammino dello sciatore sempre con maggior precisione. Andando al limite, otterremo una misura nello «spazio dei cammini» nota come misura condizionale di Wiener, condizionata dalla nostra conoscenza del fatto che lo sciatore è partito da O ed è arrivato a N.

della termodinamica statistica dei processi di non equilibrio sia (anche se meno rigorosamente, dal punto di vista matematico) fornire un completamento alla formulazione della meccanica quantistica introdotta da Richard P. Feynman mediante i cosiddetti «integrali di cammino». Nel seguito, vogliamo presentare alcune delle nuove e eccitanti prospettive di ricerca chimico-fisica che la teoria del moto browniano ha recentemente aperto.

Sebbene molti studi sulla natura fisico-matematica del moto browniano fossero stati effettuati prima del 1923, in particolare da Einstein, M. Smoluchowski, P. Langevin, J. Perrin e altri, la formulazione matematica completa di quella che è oggi nota come teoria del moto browniano fu presentata da N. Wiener, nel suo ormai famoso lavoro sugli spazi differenziali. Per questo nella letteratura matematica spesso ci si riferisce al moto browniano col nome di «processo Wiener». In quegli anni, Wiener sviluppò un'interpretazione del moto browniano come «somma su tutti i cammini», che descriveremo qualitativamente. Consideriamo una particella che subisca una serie di spostamenti, tali che l'entità e la direzione di ciascuno di essi sia indipendente dagli spostamenti precedenti. Ora, la probabilità che lo spostamento della particella browniana avvenga fra i due punti a_1 e b_1 è determinata da una funzione di distribuzione che, se la particella ha un moto simile al cammino di un ubriaco, risulta essere proprio la soluzione dell'equazione di diffusione derivata da Einstein. Il problema è determinare qual è la probabilità che, dopo n spostamenti, la particella venga a trovarsi nella zona fra a_n e b_n. Così, la probabilità dipende da un gran numero di altre grandezze aventi distribuzioni preassegnate entro un intervallo di valori. Il problema che ci troviamo di fronte è analogo a quello di uno sciatore che debba affrontare un percorso di slalom *(si veda l'illustrazione in questa pagina in basso)*. La probabilità che l'atleta passi attraverso una porta-ostacolo è data dal prodotto della larghezza della porta per la densità di probabilità, fattore peso che tiene conto sia della mobilità dello sciatore, sia della distanza fra due porte successive, sia ancora del tempo impiegato dallo sciatore per passare da una porta alla successiva. Se ora osserviamo l'atleta passare attraverso una data porta in un certo istante, potremmo desiderare di conoscere qual è la probabilità che egli passi attraverso un successivo ostacolo dopo un certo lasso di tempo. Se tale intervallo è grande, il fatto che lo sciatore passi o meno attraverso la porta non dovrebbe dipendere dal fatto che noi lo abbiamo visto passare per un'altra porta in un istante precedente. La probabilità totale sarà allora il prodotto delle probabilità individuali e possiamo immaginare che, aumentando il numero di osservazioni sullo sciatore (cioè aumentando il numero degli ostacoli sul suo percorso) e rendendo sempre più piccola la larghezza di ciascuna porta, saremmo capaci di localizzare la traiettoria dello sciatore con sempre maggiore precisione.

La difficoltà risiede nell'andare al limite di osservazioni infinitamente frequenti, e tale è appunto il problema risolto da Wiener. L'indipendenza statistica tra gli eventi può essere rigorosamente mantenuta anche nel caso limite di piccoli intervalli di tempo?

Al limite di intervalli infinitamente piccoli e di ostacoli infinitamente stretti, otteniamo quella che i matematici conoscono come «misura» di Wiener. La misura è proprio il numero che otteniamo facendo il prodotto delle varie probabilità individuali per ogni singolo evento. In più, quando non è richiesta la conoscenza del passaggio dello sciatore attraverso una qualsiasi particolare porta-ostacolo, dobbiamo sommare le probabilità su tutti i punti attraverso i quali lo sciatore potrebbe essere passato. Possiamo ora collegare il problema degli ostacoli dello slalom alla teoria delle misure, sia nella teoria classica, sia in quella quantistica.

In entrambe queste teorie il concetto di probabilità di ottenere un dato risultato in un esperimento è fondamentale. Le nozioni probabilistiche entrano nella fisica classica a causa della impossibilità di conoscere posizione e velocità di particelle in un sistema macroscopico nel quale la popolazione sia dello stesso ordine di grandezza del numero di Avogadro (circa 10^{23}). Anche se potessimo seguire ciascuna particella singolarmente, le nostre misure eseguite con strumenti macroscopici, pur non introducendo perturbazioni nel moto, potrebbero non rispecchiare il comportamento medio delle singole particelle. In meccanica quantistica, le considerazioni probabilistiche entrano per un'altra ragione: in questo caso abbiamo a che fare con particelle di dimensioni estremamente ridotte (per esempio elettroni), così che qualunque apparecchio di misura usato per rilevarne la posizione introduce una perturbazione inevitabile; le particelle che interagiscono durante la misura, per esempio i fotoni, hanno dimensioni dello stesso ordine di grandezza degli oggetti che vogliamo misurare! Classicamente potremmo chiederci qual è la probabilità P_{ab} che la misura A dia il risultato a e che, nello stesso tempo, la misura B dia il risultato b. In maniera simile, poniamo che P_{bc} sia la probabilità che la misura B dia il risultato b mentre la misura C dia il risultato c; supponiamo inoltre che P_{abc} sia la probabilità relativa al verificarsi di tutti e tre i risultati: cioè A dà a, B dà b e C dà c.

Allora, se gli eventi fra b e c sono indipendenti da quelli fra a e b (assumendo che le misure A, B, C vengano eseguite in successione temporale con lo stesso ordine), la probabilità è semplicemente il prodotto:

$$P_{abc} = P_{ab} \times P_{bc}.$$

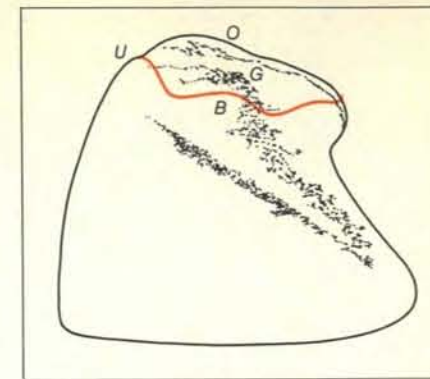
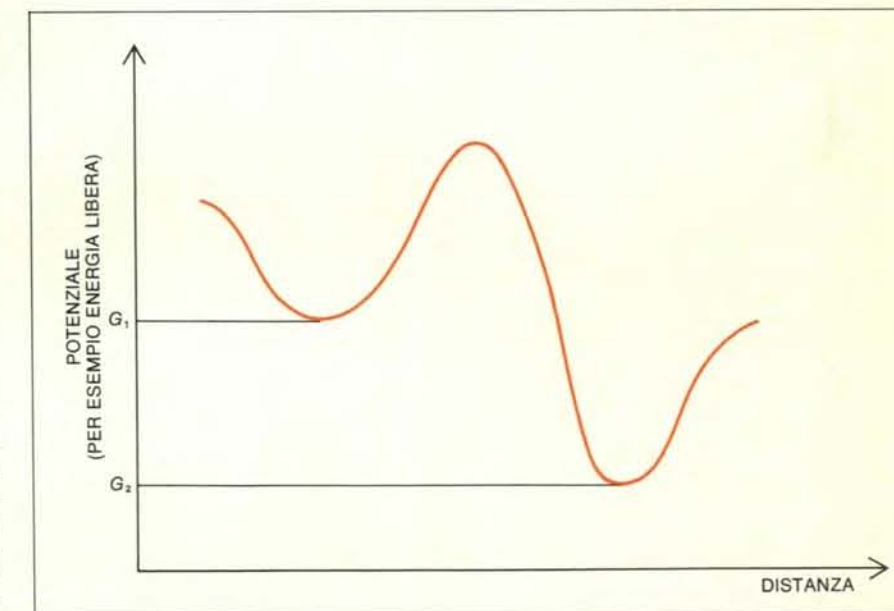
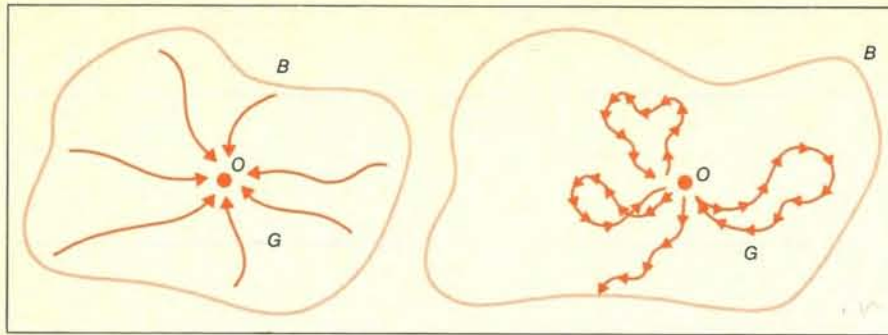
Supponiamo ora di non eseguire la misura B; la probabilità che A dia a e che C dia c è proprio:

$$P_{ac} = \text{somma su tutti i } b \ (P_{abc}),$$

questo poiché la grandezza b deve necessariamente assumere qualche valore fra le misure A e C.

Classicamente questa seconda equazione è corretta, mentre è stato dimostrato che in meccanica quantistica essa è in generale errata. Perché? Abbiamo dovuto ipotizzare che, passando da a a c, B abbia assunto qualche valore definito b. Ora, qualsiasi tentativo di misura disturberebbe il sistema in modo tale da rendere l'equazione errata per la meccanica quantistica. La perturbazione introdotta dall'apparecchio di misura sull'oggetto che deve essere misurato sarebbe sufficiente per trasformare il nostro oggetto in un sistema completamente nuovo dopo la misura! Che la seconda equazione non possa essere vera in meccanica quantistica fu affermato chiaramente per la prima volta da Werner Heisenberg col suo famoso principio di indeterminazione.

Classicamente è possibile caratterizzare un cammino con una successione di misure a istanti successivi che ci forniscano una successione di punti. Se eseguiamo un numero sufficientemente alto di misure, possiamo collegare i punti e definire una traiettoria. Sia $P_{r_1 r_2 r_3 \ldots}$ la probabilità per tale cammino. Qui r_1, r_2,... sostituiscono i risultati sperimentali a, b, c,... Se desideriamo conoscere la probabilità che r_i sia fra a_i e b_i, ecc., dobbiamo sommare (più correttamente integrare) su tutti i possibili valori intermedi. Se usassimo la soluzione dell'equazione di diffusione di Einstein come densità di probabilità, il risultato ottenuto sarebbe la misura di Wiener.

Ora, in meccanica quantistica non è possibile seguire un cammino di una particella, perché ogni volta che misuriamo la sua posizione ne disturbiamo il percorso tanto da convertirlo in un altro processo. Feynman fu il primo a capire che se si fossero sostituite le probabilità con le «ampiezze di probabilità», allora tutte le regole della probabilità classica sarebbero state valide per la meccanica quantistica. Secondo l'interpretazione ortodossa della meccanica quantistica, le ampiezze di probabilità sono grandezze complesse i cui moduli quadrati forniscono le densità di probabilità. Poiché le ampiezze di probabilità stesse non costituiscono grandezze fisicamente osservabili, è lecito intenderle quale strumento probabilistico per la descrizione dei cammini delle particelle. Tutto ciò che cambia, passando dalla visione classica di Wiener di una somma su tutti i cammini all'interpretazione di Feynman, sta nell'introduzione di una misura «complessa». Non abbiamo alcun motivo per opporci a tale assunzione perché i cammini non sono in alcun modo

Una riproduzione della superficie entropica che venne presentata a Gibbs da Maxwell nel 1875. La funzione entropia può essere rappresentata graficamente con una superficie in un sistema multidimensionale di coordinate generalizzate, con assi che indicano i valori della differenza fra le variabili termodinamiche estensive generalizzate e quelle rispettivamente corrispondenti ai valori di equilibrio. Lo stato di massima entropia O corrisponde allo stato di equilibrio. Considerando qualunque dominio arbitrario G, con frontiera B, il sistema, quasi certamente, farà la sua uscita allo stato U, che ha la massima entropia rispetto a tutti gli altri stati della frontiera, a causa delle fluttuazioni termiche che guideranno il sistema al di fuori di qualsiasi dominio circoscritto, contenente lo stato di equilibrio, nel limite in cui l'intensità del rumore termico tende a zero. Nella stessa condizione, il sistema non visiterà quegli stati le cui entropie siano minori di quella massima sulla frontiera B, a patto che esista uno stato di massima entropia in B.

Il diagramma raffigura una buca di potenziale bistabile: un campo di forze con due punti di stabilità locale. Se prendiamo in considerazione le fluttuazioni, una particella ha qualche possibilità di superare la barriera di potenziale, e così le stabilità relative delle due buche possono essere messe a confronto fra loro. Questa descrizione viene spesso usata come un modello per le reazioni chimiche, e in tal caso la distanza denota una coordinata di reazione. Le collisioni molecolari possono eccitare una molecola tanto da superare la barriera, così che si verifichino la rottura del legame molecolare e la dissociazione della molecola. Le collisioni casuali sono descritte da un «rumore bianco» la cui intensità determina la temperatura dei reagenti. Una volta superata la barriera di potenziale, c'è la possibilità che si vengano a formare nuovi e diversi legami molecolari, e che di conseguenza il sistema vada a cadere nello stato di equilibrio più stabile, restandovi in quiete (a meno delle fluttuazioni termiche). Il tasso con cui le collisioni molecolari spingono le molecole al di là della soglia della buca di potenziale è determinato dalla cinetica della reazione.

A sinistra le linee di flusso per il caso in cui lo stato stazionario *O*, nel dominio *G*, è stabile. Le fluttuazioni termiche casuali possono essere immaginate come causa di una lenta diffusione del sistema in un campo di flusso deterministico. Poiché lo stato stazionario è stabile, il sistema deve diffondere «contro il flusso» per raggiungere il contorno *B*. A causa delle fluttuazioni termiche casuali, il sistema può passare da uno stato stazionario stabile ad un altro, ma esso deve necessariamente andare contro il flusso, almeno per una parte della traiettoria. A destra, il moto del sistema, precedente alla sua uscita dal contorno *B*, a partire da un intorno dello stato stazionario stabile *O* appartenente a un dominio limitato *G*, ha la seguente natura: il sistema è attratto dallo stato stazionario, qualunque sia lo stato in *G* in cui esso si trovi, e il moto verso lo stato stazionario è descritto da un'equazione deterministica cinetica. Giunto nelle vicinanze di *O*, il sistema può effettuare escursioni in quegli stati che sono più distanti, per poi tornare indietro attratto dallo stato stazionario. Se accade che il sistema raggiunge il contorno, allora il suo moto, in precedenza, non sarà lento e continuo, ma piuttosto avrà l'aspetto di un moto a balzi in cui percorre una distanza definita in un tempo definito: altrimenti verrebbe risucchiato verso lo stato stazionario.

osservabili. Il rigore matematico di tale formulazione è ancora in fase di perfezionamento poiché insorgono problemi di convergenza, e in più non è stata individuata la provenienza del rumore che impedisce una trattazione deterministica della meccanica quantistica.

Wiener mise in relazione la densità di probabilità con il processo di diffusione di una particella «libera», mentre Feynman diede, in forma assiomatica, regole per il calcolo della densità di ampiezza di probabilità, dimostrando a posteriori che essa soddisfa l'equazione di diffusione più notevole della meccanica quantistica: l'equazione di Schrödinger. Ha senso allora chiedersi a quale processo fisico queste regole formali corrisponda-no. Per rispondere a questa domanda dobbiamo introdurre una linea di ricerca parallela allo studio delle equazioni di diffusione: la teoria delle equazioni differenziali stocastiche. P. Langevin può essere considerato il fondatore di tale formalismo che consente una rappresentazione semplice (sebbene matematicamente complessa) del processo fisico chiamato moto browniano.

Caratteristica di questa interpretazione è il fatto che le equazioni del moto utilizzate nell'analisi dei sistemi fisici sono una riformulazione della legge del moto di Newton: $F = ma$, cioè forza = massa × accelerazione, con l'eventuale inclusione delle forze di attrito, o dissipative. Sebbene l'applicazione diretta di $F = ma$ conduca a equazioni differenziali del secondo ordine, è sempre possibile, mediante l'introduzione di variabili aggiuntive, trasformare queste equazioni in un sistema di equazioni differenziali accoppiate del primo ordine (ma sovente non lineari), aventi la forma:

$$\frac{dx(t)}{dt} = b[x(t)]$$

dove $x(t)$ è in generale un vettore di dimensione $n>1$, detto «stato del sistema» al tempo t. Se si vogliono prendere in considerazione anche fluttuazioni del sistema simili al moto perpetuo delle particelle browniane, mantenute in tale stato dalle collisioni casuali con le molecole più leggere del fluido circostante, occorre aggiungere, come di consueto, una forza casuale o fluttuante al membro di destra di quest'ultima equazione; ciò provoca la conversione dell'equazione differenziale deterministica in un'altra, nota come equazione differenziale stocastica (o equazione di Langevin), avente la forma:

$$\frac{dx(t)}{dt} = b[x(t)] + f(t).$$

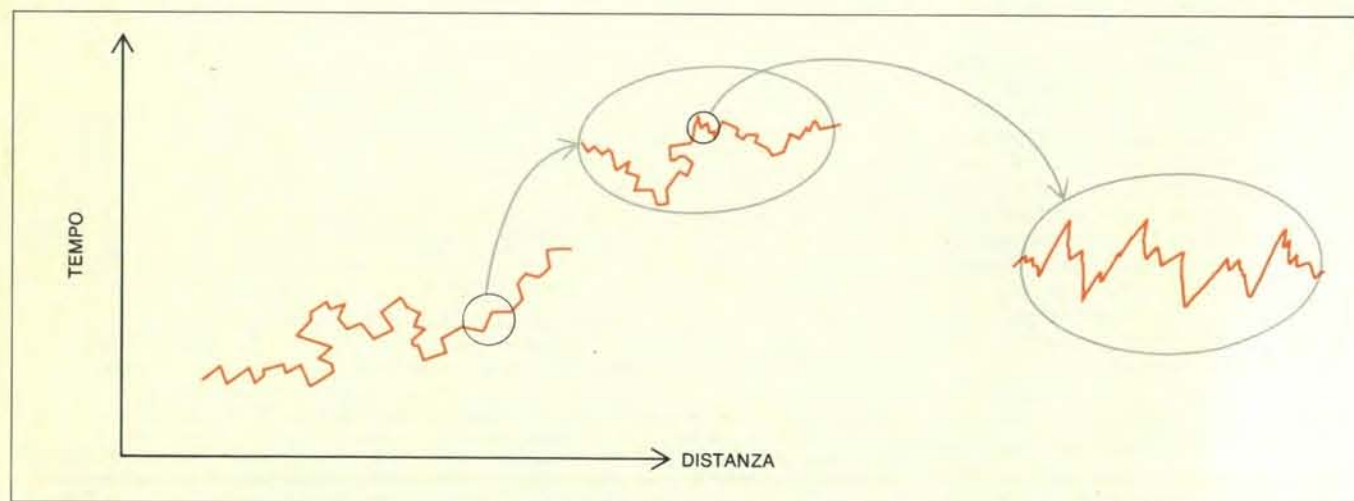
Le equazioni differenziali del tipo dell'equazione di Langevin connettono due mondi separati: il mondo macroscopico rappresentato dal «vettore di deriva» b, e il mondo microscopico, rappresentato dalla forza fluttuante f. Nella formulazione originale di Langevin, $x(t)$ rappresentava la quantità di moto della particella browniana e $b[x(t)]$ l'attrito dinamico agente sulla particella; la parte fluttuante $f(t)$ caratterizzava il moto browniano. Langevin arguì che tale decomposizione doveva essere valida in quanto il moto ha luogo su due scale di tempo largamente separate fra loro: una corta, secondo la quale varia rapidamente la forza fluttuante $f(t)$ (ricordiamo che una particella subisce normalmente circa 10^{21} collisioni al secondo) e una lunga, rispetto alla quale si manifestano gli effetti dell'attrito dinamico.

Risolvere un'equazione differenziale stocastica del tipo di quella di Langevin non è come risolvere un'equazione differenziale ordinaria: infatti l'equazione di Langevin mette in gioco una forza fluttuante $f(t)$ che ha proprietà definite solo statisticamente. In assenza di una conoscenza specifica delle forze casuali, si assume comunemente che $f(t)$ sia un cosiddetto processo casuale gaussiano di «rumore bianco»: il termine «rumore bianco» è usato per analogia con la luce «bianca», poiché in entrambi i casi lo spettro è costante, ovvero di ampiezza indipendente dalla frequenza. Si deve tuttavia considerare che un vero rumore bianco non può esistere nel mondo reale: qualsiasi rumore osservabile, infatti, indipendentemente da quanto risulti «piatto» il suo spettro alle basse frequenze, si annulla regolarmente alle alte frequenze. Il modello del rumore bianco, quindi, porta inevitabilmente alla cosiddetta «catastrofe ultravioletta», conseguenza questa che vanificò i tentativi di trattare lo spettro del «corpo nero» sulla base del principio di equipartizione dell'energia. Perciò si tratta di una idealizzazione matematica che può soltanto approssimare la realtà.

In effetti esistono altre forme di rumore; per esempio il rumore granulare creato per emissione spontanea dal catodo di un elettrone che, raggiungendo l'anodo, produce una corrente. Analogamente l'emissione di altri elettroni è descritta da una distribuzione casuale di tempi di emissione che statisticamente ha l'aspetto di una distribuzione di Poisson. Comunque, se applichiamo la legge dei grandi numeri, sappiamo che al limite di molti eventi indipendenti le distribuzioni statistiche tendono sempre a una distribuzione gaussiana, e ciò ci riconduce al rumore bianco che è sempre stazionario nel tempo e ha uno spettro uniforme.

A questo punto viene da chiedersi: quale delle due descrizioni, deterministica (equazioni differenziali del primo ordine) o statistica (equazioni differenziali stocastiche di Langevin), si avvicina di più a una descrizione realistica di ciò che accade nei sistemi dinamici? Prendiamo in considerazione, per esempio, la termodinamica classica: essa prevede che le fluttuazioni relative di una variabile termodinamica estensiva siano proporzionali all'inverso della radice quadrata del numero delle particelle. Al limite termodinamico, in cui il numero delle particelle e il volume tendono all'infinito in modo tale che il loro rapporto rimanga costante, le fluttuazioni relative tendono a zero, e la distribuzione si raccoglie sempre più attorno al valore atteso, che è quello che la termodinamica prevede essere il valore sperimentale. Ma noi sappiamo che si verificano sempre piccole deviazioni dalle equazioni di stato termodinamiche: se due sistemi sono preparati in modo identico, non è detto che successivamente si comporteranno esattamente allo stesso modo.

È questo che intendiamo quando sosteniamo la necessità di prendere in considerazione le fluttuazioni. Inoltre, in piccole regioni di spazio, oppure in prossimità di un punto critico, le proprietà di un sistema fluttuano largamente intorno ai valori previsti dalle equazioni di stato deterministiche, e il comportamento macroscopico futuro è determinato dalle fluttuazioni iniziali. Consideriamo inoltre il caso di una buca di potenziale bistabile, del tipo di quelle comunemente usate per descrivere le reazioni chimiche *(si veda l'illustrazione a pagina 000)*. La termodinamica classica prevede che il sistema tenderà a trovarsi sempre in corrispondenza del minimo più basso (G_2) perché esso corrisponde all'energia libera più bassa; un sistema che si trovasse però nel minimo più alto (G_1) sarebbe destinato a rimanervi, in assenza di fluttuazioni: quindi la previsione della termodinamica classica non può essere del tutto legittima. In altri termini non esiste per il sistema alcun modo per accorgersi della presenza di un minimo dell'energia libera termodinamicamente più stabile.

È ben noto che le fluttuazioni giocano un ruolo cruciale nel caso di sistemi fisici che diventano instabili (in modo appros-

I cammini tipici di una particella browniana sono fortemente irregolari se osservati su scala molto fine, come mostrato qui. Così, è possibile definire una velocità media, ma non esiste una velocità istantanea. Questo significa che i cammini del moto browniano non sono differenziabili.

Mentre l'entropia del sistema tende ad aumentare fino a raggiungere il valore massimo all'equilibrio, l'entropia generata dalle correlazioni statistiche fra stati di non equilibrio tende, in media, a diminuire col trascorrere del tempo. L'entropia addizionale, prodotta dalle correlazioni statistiche, distrugge la proprietà termodinamica di additività, che torna a valere solo per tempi lunghi, ove le correlazioni statistiche abbiano avuto abbastanza tempo per esaurirsi.

Schematizzazione delle fluttuazioni intorno al cammino medio per la regressione del sistema fino all'equilibrio. Poiché le medie e i valori più probabili di una distribuzione gaussiana sono identici, il cammino medio coinciderà con il cammino più probabile per la regressione di una fluttuazione solo nel caso che le fluttuazioni siano gaussiane. Per fluttuazioni non-gaussiane, solo nel limite di intensità di rumore termico molto piccola, le distribuzioni delle variabili termodinamiche fluttuanti diverranno così nette che il comportamento medio e quello più probabile torneranno a coincidere.

simativo possiamo dire che un sistema dinamico è stabile se è insensibile a piccole perturbazioni). Comunque le nostre equazioni, sia quella deterministica, sia quella di Langevin, prevedono un comportamento qualitativamente differente. Assumiamo che esista uno stato deterministicamente stabile, isolato e stazionario, determinato dalla scomparsa della deriva: $b(O) = 0$ in cui con O si simboleggia lo stato stazionario. Questo significa che il sistema non sta «evolvendo». Allora ambedue le equazioni saranno concordi nel prevedere che il sistema tenderà inizialmente ad agire in modo da eliminare ogni eventuale variazione del suo stato causata da perturbazioni, ristabilendo così lo stato stazionario, O.

Ambedue le equazioni sono concordi nel dire che il sistema «scivolerà» lungo una curva che sia soluzione del sistema deterministico. Ora l'equazione differenziale del primo ordine prevede che il sistema si avvicinerà asintoticamente allo stato stazionario, e null'altro. Al contrario l'equazione di Langevin prevede che il sistema si avvicinerà a un piccolo intorno di O, nel quale trascorrerà la maggior parte del suo tempo. Ma, tenendo conto della forza casuale, il sistema avrà sempre una probabilità di saltare in un altro stato in G. Il moto non è continuo; consiste invece di salti discontinui, poiché il sistema avverte sempre l'attrazione dello stato stazionario. Presto o tardi, il sistema raggiungerà la frontiera B con una probabilità prossima a 1. Certamente il tempo che impiegherà per raggiungere tale frontiera sarà più lungo di quello impiegato in tentativi infruttuosi di raggiungere B. Comunque l'equazione di Langevin dice che il sistema abbandonerà quasi certamente qualsiasi dominio circoscritto in un tempo finito. Il sistema è sempre instabile, indipendente dalla grandezza della forza fluttuante, posto che tale forza agisca anche nello stato stazionario che dall'equazione differenziale ordinaria del primo ordine risulta essere globalmente stabile (la forza fluttuante è in media zero).

Una simulazione col metodo di Monte Carlo di un sistema dinamico bidimensionale lineare, soggetto a piccole eccitazioni del tipo del rumore bianco, descriventi le fluttuazioni termiche, è stata eseguita da R. G. Williams, che ne ha riferito sul «SIAM Journal on Applied Mathematics» nel 1981. Lo stato stazionario è posto all'origine, e la figura mostra i risultati per cinque traiettorie di fuga. Sebbene il sistema trascorra la maggior parte del tempo nelle immediate vicinanze dello stato stazionario, effettua anche deviazioni erratiche, allontanandosene. Il moto di diffusione contrario al flusso non è un processo lento; copre distanze finite in intervalli di tempo relativamente corti. Se il sistema non raggiunge la frontiera circolare, viene attratto dalle vicinanze dello stato stazionario, dove inizierà un'altra escursione a un tempo successivo. Il risultato conferma la previsione teorica che le traiettorie si stringono attorno alla traiettoria a tempo rovesciato del sistema deterministico, che collega lo stato stazionario con lo stato della frontiera più vicino avente massima entropia, se si diminuisce sufficientemente l'intensità del rumore.

Il problema che abbiamo descritto è quello di una particella che diffonde in direzione contraria al flusso. Questo problema diviene fisicamente più interessante se sono presenti più stati stazionari del moto. In tal caso esso può rappresentare la fuga di una particella da una buca di potenziale divenendo così un modello per le reazioni chimiche, per la diffusione nei cristalli, per le transizioni nelle giunzioni Josephson, così come per l'attraversamento delle barriere di potenziale da parte di particelle quantomeccaniche (effetto «tunnel»).

La presenza di fluttuazioni termiche casuali introduce un limite superiore alla precisione con la quale è possibile specificare un dato stato macroscopico di un sistema. Agli inizi degli anni cinquanta cominciarono intense ricerche per lo studio delle equazioni differenziali stocastiche del tipo di quella di Langevin; ne è risultato un nuovo tipo di calcolo, che prende il nome dal suo ideatore: Itô. Il calcolo stocastico di K. Itô è basato sull'osservazione che il cammino di una particella browniana è molto irregolare, se osservato con sufficiente dettaglio. (Abbiamo trascurato questo particolare nel formulare l'equazione di Langevin, che rimane comunque formale.) Ne consegue che, nel moto browniano non lo spostamento Δx, ma il suo quadrato è proporzionale a Δt, il coefficiente di proporzionalità essendo proprio due volte la costante di diffusione, D. In più è noto, fino dal 1933, che il moto browniano dà luogo alla relazione di indeterminazione:

$$\Delta x^2 \geqslant 2D\Delta t$$

in diretta analogia con la relazione di indeterminazione vigente nella meccanica quantistica:

$$\Delta x^2 \geqslant \frac{h}{m}\Delta t$$

in cui h è la costante di Planck e m è la massa della particella. Si può quindi osservare che in meccanica quantistica il rapporto $h/2m$ svolge lo stesso ruolo del coefficiente D nella teoria dei processi stocastici, e il calcolo stocastico può essere applicato in modo formale alla meccanica quantistica, sebbene non vi sia alcun processo di diffusione reale. La relazione di indeterminazione del moto browniano sta a significare che dobbiamo modificare le usuali regole di differenziazione (o di integrazione). Per ottenere il differenziale di una funzione è ora necessario prendere due termini nello sviluppo in serie di Taylor e sostituire il termine quadratico Δx^2 con il suo valor medio $2D\Delta t$.

La relazione di indeterminazione del moto browniano è una manifestazione delle correlazioni statistiche fra stati di non equilibrio attraverso i quali il sistema passa a istanti di tempo successivi.

La presenza di fluttuazioni termiche rende necessaria la transizione verso un'interpretazione probabilistica. La termodinamica statistica offre una connessione naturale fra la probabilità di uno stato e la sua entropia. Perciò l'entropia, anziché l'energia libera, ha un ruolo privilegiato quando vengono introdotti concetti probabilistici in termodinamica. Boltzmann fu il primo a riconoscere la connessione fra probabilità e entropia. Sfortunatamente visse in un'epoca in cui il determinismo della meccanica classica pervadeva le scienze naturali e soltanto dopo la sua morte Einstein riprese le sue idee e le applicò con tanto successo al moto browniano. Questo avvenimento, insieme alla spiegazione di Max Planck della radiazione del corpo nero, annunciò la nascita dell'era atomica.

Einstein mise in relazione la densità di probabilità $P(x)$, per una fluttuazione spontanea in uno stato di non equilibrio, con la diminuzione dell'entropia ΔS secondo la relazione:

$$P(x) \propto \exp [\, \Delta S(x) \,].$$

Questo implica che la probabilità, per differenti stati di non equilibrio, è proprio il prodotto delle probabilità individuali, confermando la proprietà di additività dell'entropia. Possiamo apprezzare la formula di Einstein per una fluttuazione spontanea dall'equilibrio come una forma limite, valida per tempi lunghi, qualora, cioè, le correlazioni statistiche abbiano avuto il tempo di attenuarsi, o al limite di piccole fluttuazione termiche. In generale vi sarà un'entropia addizionale generata dalle correlazioni statistiche fra stati di non equilibrio. Gli andamenti medi delle entropie di stato, opposti a quelli delle entropie correlazionali per stati di non equilibrio, sono schematizzati nella figura a pagina 29 in alto. Si può vedere che la presenza di correlazioni statistiche distrugge le proprietà termodinamiche di additività che caratterizzano l'equilibrio e che alla formula di Einstein per la densità di probabilità bisogna aggiungere una densità supplementare che tenga conto delle probabilità di transizione fra stati di non equilibrio.

Noi abbiamo derivato un'espressione generale per la densità di probabilità di transizione $P(x \rightarrow y)$, relativa a due stati di non equilibrio x e y; essa è data dall'analogo cinetico della formula di Einstein:

$$P(x \rightarrow y) \propto \exp [\, {}^1\!/_2 \,(\Sigma + \Delta S) \,]$$

in cui Σ è l'entropia congiunta e ΔS è la differenza di entropia fra i due stati di non equilibrio. L'entropia congiunta ha la proprietà di ridursi, dopo lungo tempo, alla somma delle entropie, e quest'ultima formula si riduce alla formula di Einstein.

Il principio generale della termodinamica del non equilibrio che governa l'evoluzione dei processi irreversibili verso l'equilibrio è:

$$\frac{d\overline{\Sigma}}{dt} \leqslant O$$

in cui la sopralineatura denota l'operazione di media. La diseguaglianza afferma che le correlazioni statistiche fra stati di non equilibrio sono, in media, una funzione decrescente del tempo. Usando le parole di Lars Onsager, essa esprime il fatto che tutti i sistemi fisici tendono a dimenticare il loro passato. Infine, la diseguaglianza può essere considerata come l'analogo stocastico del celebre teorema H di Boltzmann. Essa fornisce un meccanismo fisico per la conversione del nostro analogo cinetico nella formula di Einstein, al limite dei tempi lunghi.

Si può comprendere la differenza fra il modello deterministico (equazione differenziale del primo ordine) e quello stocastico (equazione di Langevin) dei sistemi dinamici considerando che l'entropia è una funzione della variabile x che è lo scarto dall'equilibrio di una variabile di stato. Utilizzando l'equazione differenziale deterministica si scopre che l'entropia tenderebbe a zero per tempi lunghi, poiché per definizione x si annulla all'equilibrio (tempi lunghi). Se viene invece usata l'equazione di Langevin per calcolare l'entropia, troviamo che essa tende alla sua corretta forma d'equilibrio, al limite per tempi lunghi; questo dimostra che la presenza di fluttuazioni termiche casuali è fondamentale per lo stabilirsi della distribuzione d'equilibrio. In altre parole, l'equilibrio macroscopico corrisponde a uno stato medio attorno al quale il sistema fluttua a causa delle fluttuazioni termiche casuali. Infatti, mediando l'equazione di Langevin otteniamo l'equazione macroscopica fenomenologica della termodinamica dei processi irreversibili. Questa affermazione è stata enunciata per la prima volta da Onsager nella sua ipotesi della regressione delle fluttuazioni: la regressione delle fluttuazioni di non equilibrio obbedisce, in media, alle leggi fenomenologiche della termodinamica dei processi irreversibili. Il cammino che coincide con la soluzione deterministica della equazione cinetica (equazione differenziale del primo ordine) è chiamato «cammino termodinamico»; le fluttuazioni intorno al cammino termodinamico sono illustrate a pagina 29 in basso.

Il sistema progredisce verso un piccolo intorno dello stato di equilibrio caratterizzato dalla distribuzione statistica data dalla formula di Einstein; qui il sistema trascorre la maggior parte del suo tempo. Ora spostiamo il nostro asse del tempo a qualche istante lontano nel passato, affermando che il sistema invecchiato deve pur essere stato all'equilibrio molto tempo fa. A causa delle fluttuazioni termiche, il sistema avrà pure compiuto escursioni in stati a entropia inferiore di quella dello stato di equilibrio ma, con la diminuzione dell'intensità del rumore termico, il sistema evita gli stati nel dominio G con un'entropia inferiore di quello stato della frontiera B che possiede l'entropia massima, e attraverso il quale dovrebbe aver luogo l'uscita dal dominio G. In altre parole, con la diminuzione dell'intensità del rumore, il sistema non avrà energia sufficiente per visitare quella parte del dominio in cui l'entropia è inferiore a questo massimo sulla frontiera.

Nel caso di piccole fluttuazioni termiche, la prima uscita dalla frontiera avrà luogo, quasi certamente, in prossimità di quello stato U su B che rende massima l'entropia congiunta, Σ, soggetta alla condizione che il sistema sia stato nella vicinanza dello stato di equilibrio molto tempo fa. La fuga sarà avvenuta entro una piccola zona cilindrica, come un tubicino, attorno a un percorso che rende massima l'entropia congiunta. Sotto le stesse condizioni, questo cammino è l'immagine speculare nel tempo del percorso termodinamico, o deterministico, per la regressione delle fluttuazioni. Pertanto, il processo manifesta «una simmetria tra passato e futuro».

La linea piena mostra l'avvicinamento, entro un intorno Δ dello stato di equilibrio *O*, lungo il percorso più probabile o cammino termodinamico. Δ è una misura della grandezza delle fluttuazioni termiche attorno allo stato di equilibrio. Il cammino più probabile per l'uscita dalla regione *G* sarà entro un piccolo tubicino attorno al cammino che sia l'immagine speculare nel tempo del cammino più probabile per la regressione di una fluttuazione (*linea tratteggiata*), al limite per l'intensità del rumore termico tendente a zero.

Questo comportamento è stato verificato con esperimenti di simulazione al calcolatore col metodo di Monte Carlo. Al diminuire dell'intensità del rumore termico, gli ultimi segmenti dei cammini che conducono al contorno si concentrano, quasi certamente, intorno a quella traiettoria che è l'immagine speculare del cammino più probabile per la regressione di una fluttuazione. Nella illustrazione a pagina 32 viene mostrato il processo chimico di Lotka, nel quale la rappresentazione delle curve integrali nel piano di fase indica che qualunque perturbazione del sistema «dinamico» causerà una reazione diretta alla restaurazione dell'equilibrio, che avrà luogo asintoticamente nel tempo, per piccole perturbazioni. A causa delle fluttuazioni termiche, è certo che il processo potrà sfuggire da qualunque dominio limitato contenente lo stato di equilibrio stabile. Per intensità molto piccole del rumore termico, la traiettoria di fuga convergerà con l'immagine speculare della traiettoria per il cammino più probabile della regressione di una fluttuazione.

Abbiamo chiamato l'immagine speculare del percorso termodinamico nel tempo col nome di «cammino antitermodinamico». Sebbene, a partire da un intorno dello stato di equilibrio, l'uscita dalla frontiera lungo il cammino antitermodinamico sia improbabile, anche l'uscita lungo qualsiasi altro percorso sarà improbabile, e in misura maggiore. Poiché, però, il tempo che il sistema trascorre nelle vicinanze dello stato di equilibrio è illimitato, l'uscita meno improbabile fino alla frontiera avrà luogo, prima o poi, e naturalmente nel modo più probabile. Ovviamente, il sistema avrebbe potuto raggiungere la frontiera lungo un'infinità

Lo schema della reazione chimica di Lotka viene usato anche come modello per fenomeni oscillatori che possono verificarsi negli ecosistemi. Il substrato iniziale *A* e il prodotto finale *P* sono mantenuti in quantità costanti dai flussi esterni. Gli intermedi *X* e *Y* variano nel tempo, in un modo descritto da equazioni deterministiche. Oscillazioni di queste concentrazioni nel tempo si verificano in situazioni lontane dall'equilibrio chimico, dove le reazioni inverse sono trascurabili in confronto alle reazioni dirette. In questo caso lo stato stazionario non è più stabile, nel senso che il sistema non agisce contro le perturbazioni che spostano il sistema dallo stato stazionario. Un fatto assai interessante è che anche quando ci troviamo nelle vicinanze dello stato di equilibrio, in cui si applica la legge dell'azione di massa, lo stato di equilibrio (situato all'origine del piano [X-X_O, Y-Y_O]) è instabile in presenza delle fluttuazioni termiche. Le equazioni cinetiche impiegate nella descrizione della reazione chimica sono ora del tipo dell'equazione di Langevin e, diversamente dal caso deterministico, il sistema compirà la sua uscita fuori da qualunque regione limitata contenente lo stato di equilibrio. Per piccole deviazioni dallo stato di equilibrio, la superficie entropica è convessa e ha il massimo di entropia S_O in corrispondenza dello stato di equilibrio. Le curve integrali, che sono ovunque tangenti al moto del sistema, sono mostrate in colore chiaro. Esse descrivono il moto del sistema nel suo avvicinarsi allo stato di equilibrio. In presenza di fluttuazioni termiche, il sistema farà la sua uscita da qualsiasi regione *G*, con contorno *B*, racchiudente lo stato di equilibrio. Nel limite di disturbi di intensità molto piccole, il cammino più probabile che il sistema seguirà nell'effettuare la sua uscita da *G* è la traiettoria del sistema deterministico, con il tempo rovesciato, che collega lo stato di equilibrio *O* con lo stato di massima entropia S^* sul contorno *B*, a patto che esista uno stato di massima entropia in *B*. Questa curva integrale col tempo invertito è rappresentata in colore più intenso e l'uscita si verificherà, quasi certamente, dallo stato *U* della frontiera che ha la massima entropia, al limite per il rumore termico tendente a zero. La traiettoria più probabile per l'uscita può essere trovata dalla condizione di stazionarietà dell'entropia congiunta, soggetta alla condizione che il sistema si sia trovato nello stato di equilibrio molto tempo prima.

di altri cammini, diversi da quello antitermodinamico. La probabilità del passaggio attraverso un tubicino, contenente uno di questi cammini qualsiasi, è infinitamente piccola se paragonata alla probabilità che il cammino avvenga all'interno di un tubicino contenente il percorso antitermodinamico. Occorre tuttavia notare che la somma delle probabilità di passaggio lungo gli infiniti tubicini possibili può anche essere più grande di quest'ultima.

Non è un caso che la meccanica quantistica ci dia una relazione di indeterminazione formalmente identica alla relazione di indeterminazione del moto browniano. Sin dalle prime formulazioni, si era intuito che la meccanica quantistica era incompleta: se fosse stato possibile specificare ulteriori variabili per descrivere qualche meccanismo interno fondamentale, ciò che ora è soltanto probabile sarebbe divenuto «certezza», rivelando così un determinismo soggiacente. Sebbene vi sia molta letteratura sulle teorie delle «variabili nascoste», nessuna di queste ha avuto successo nel dare un'interpretazione completamente soddisfacente di tutta la meccanica quantistica. Le formulazioni stocastiche della meccanica quantistica sono una via di mezzo fra la meccanica quantistica tradizionale e la teoria delle variabili nascoste, ipotizzando interazioni casuali fra le particelle quantistiche e il mezzo ipotetico nel quale esse si muovono. L'idea che una particella quantistica possa essere soggetta a un moto browniano classico è molto suggestiva, anche se prima sarà necessario riconciliare quest'idea col fatto che le traiettorie delle particelle quantistiche non sono osservabili, poiché qualsiasi tentativo di osservare la particella quantistica richiede un'interazione che perturba il sistema osservato. A questo punto ci piacerebbe credere che le probabilità siano reali, positive e normalizzabili, e che prevedano la frequenza con cui si verificano gli eventi reali. Per esempio, potremmo assegnare una distribuzione di probabilità a tutte le traiettorie possibili di una particella browniana. In meccanica quantistica, però, non vi è motivo per supporre l'esistenza di una distribuzione positiva e reale di probabilità per una data traiettoria, non potendosi eseguire più di una misura sullo stesso processo. (Una singola misura sulla particella causa la realizzazione di un evento osservabile, per il quale otteniamo una densità di probabilità reale che è sia positiva sia normalizzata.)

A differenza della meccanica classica, la meccanica quantistica è selettiva, nel senso che fa distinzione fra ciò che è fisicamente osservabile e ciò che è matematicamente «misurabile». Così la meccanica quantistica stocastica associa un processo, descritto da un'equazione del tipo di quella di Langevin, a ogni stato quantistico dinamico, e tutte le medie sono eseguite con una misura complessa della probabilità. La teoria è di per sé selettiva poiché risultano fisicamente osservabili solo quantità medie che siano reali. L'influenza dei campi esterni sul sistema entra attraverso la definizione del moto di deriva, e la formulazione usuale della meccanica quantistica mette in relazione il campo esterno con la funzione d'onda. In questo senso si può dire che la funzione d'onda determina lo stato del sistema, poiché specifica il moto tramite la deriva. Quindi non v'è alcuna differenza fisica fra il moto classico browniano e la meccanica quantistica; la sola differenza è di ordine matematico: l'uso di misure di probabilità complesse separa le osservazioni che hanno probabilità reali e positive di verificarsi da quelle non osservabili, che risultano poi essere quantità complesse. È un merito e un vantaggio dell'impostazione stocastica, che tutti i concetti probabilistici vengano introdotti in modo completamente classico. Se non altro il moto browniano ha fornito un modo per interpretare e capire da un punto di vista fisico cose note da tempo, ma astratte.

Immagini radar della Terra dallo spazio

Sistemi radar orbitanti, in grado di sintetizzare immagini in base agli echi riflessi di segnali a microonde, forniscono nuove informazioni sulle caratteristiche della superficie della Terra

di Charles Elachi

Quasi tutte le immagini della Terra riprese dallo spazio sono state ottenute con procedimenti molto simili alla normale fotografia: raccolta, messa a fuoco e registrazione della radiazione solare riflessa. Altre tecniche di telerilevamento sono basate sulle radiazioni termiche o su quelle a microonde emesse dalla superficie terrestre. Da poco è stata però introdotta una tecnica sostanzialmente differente per i rilevamenti mediante satelliti. Un fascio di microonde viene diretto obliquamente verso la Terra da un sistema radar in orbita, il cui ricevitore rileva la radiazione riflessa dalla quale è possibile ottenere un'immagine sintetizzata. I sistemi passivi di ripresa ad alta definizione lavorano in generale in luce visibile o nelle adiacenti regioni infrarossa e ultravioletta dello spettro elettromagnetico. Un radar attivo su satellite consente di effettuare lo studio dettagliato della superficie terrestre nella regione delle microonde, raccogliendo informazioni su proprietà che finora non erano rilevabili.

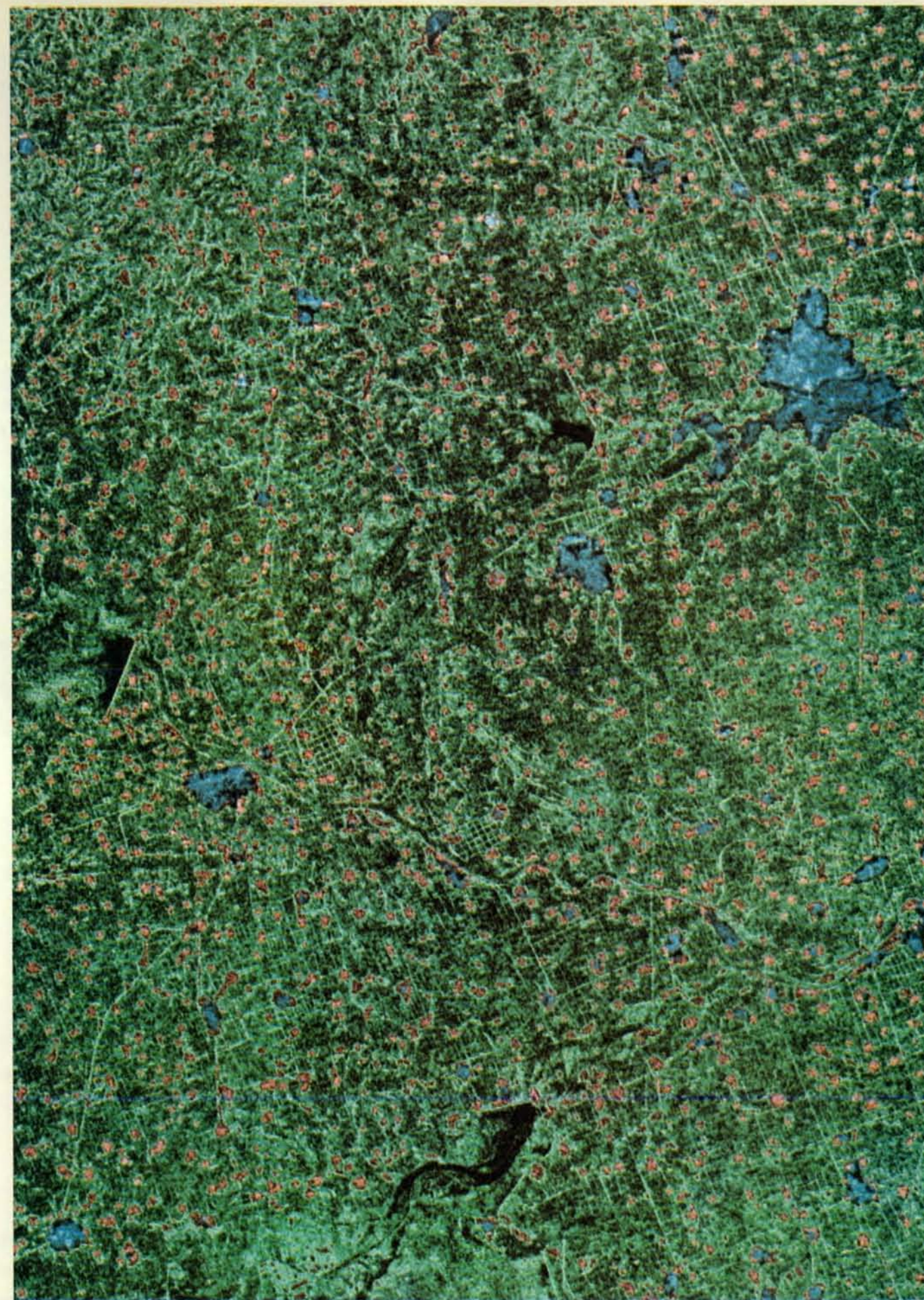
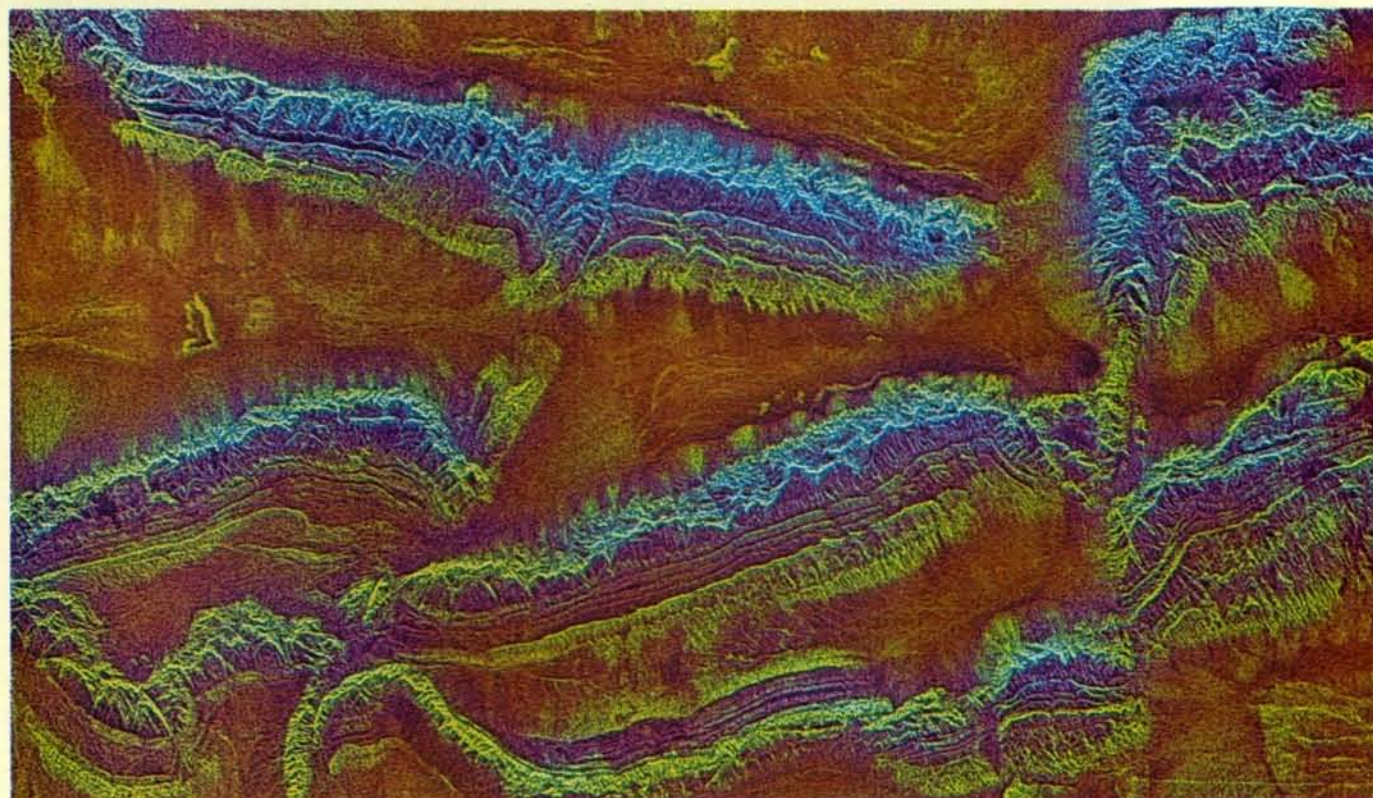
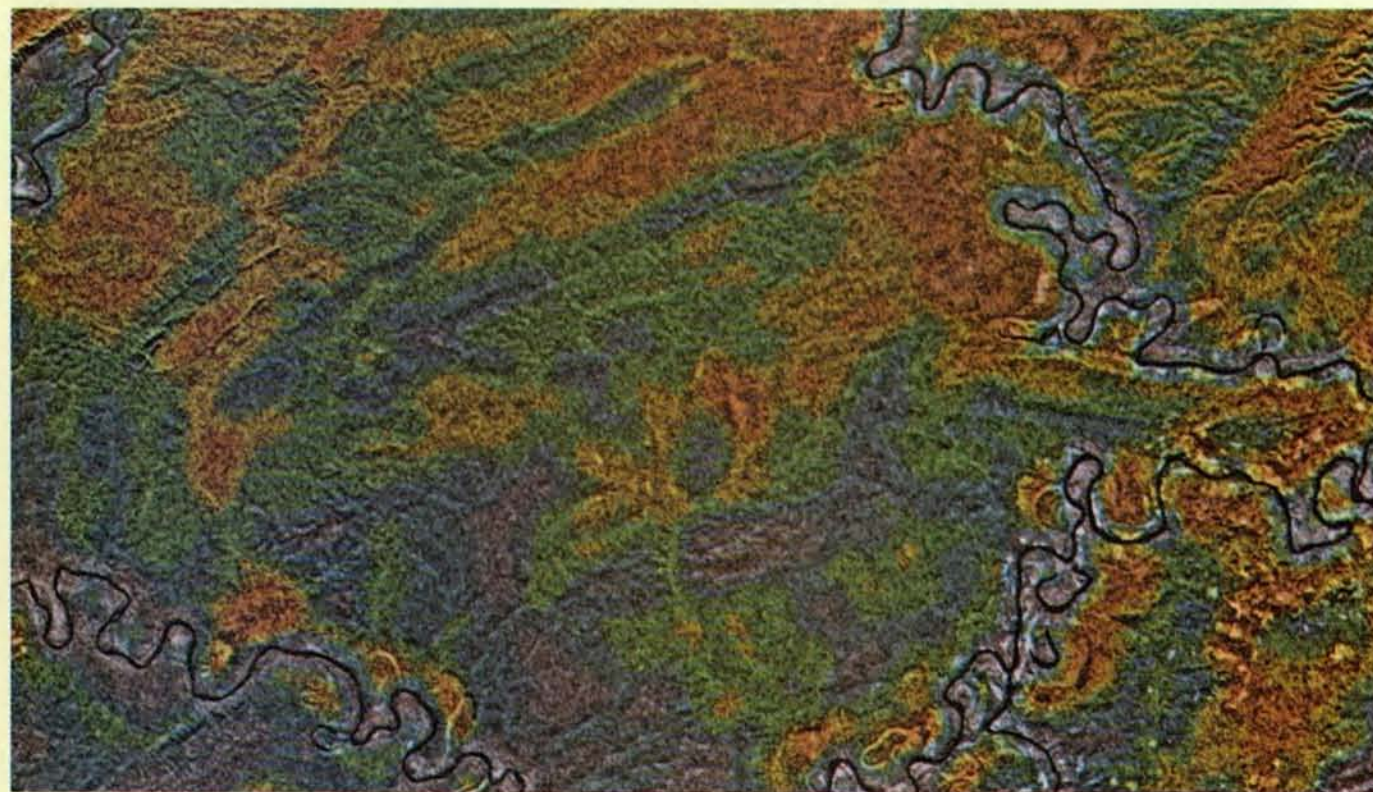
Un sistema radar attivo, poiché possiede una sorgente propria di illuminazione, non dipende dalla luce solare e può quindi funzionare in qualunque momento del giorno e della notte. Inoltre, poiché nubi, nebbia e precipitazioni hanno ben scarsa influenza sulle microonde, un sistema del genere può essere impiegato in qualunque condizione meteorologica. Questa capacità di fornire immagini con continuità è fondamentale per l'osservazione di fenomeni dinamici, per esempio delle correnti oceaniche, dello spostamento di banchi di ghiaccio galleggianti o di caratteristiche della vegetazione in fase di cambiamento.

La luminosità di un determinato punto dell'immagine radar dipende dall'intensità dell'energia a microonde riflessa dal punto corrispondente sulla superficie della Terra, che viene rilevata dal ricevitore radar. L'intensità della riflessione dipende a sua volta sia dalle proprietà fisiche della superficie (quali l'inclinazione rispetto al raggio incidente, la scabrosità e la copertura vegetale), sia da quelle elettriche (soprattutto la conduttività, che dipende da vari fattori fra cui la porosità e il contenuto di acqua del suolo). Nelle regioni del visibile e dell'infrarosso vicino, il potere riflettente della Terra è determinato in prevalenza dalla composizione chimica del terreno, mentre a lunghezze d'onda maggiori, cioè nell'infrarosso termico, esso è sostanzialmente funzione della sola capacità termica del suolo. In breve, occorre una combinazione di sensori per una descrizione completa della superficie terrestre.

Finora sono stati posti in orbita intorno alla Terra due sistemi radar per immagini. Il più recente era installato a bordo della navetta spaziale *Columbia* durante il suo secondo volo nel novembre 1981: denominato *SIR-A (Shuttle Imaging Radar -A)*, ha fornito riproduzioni straordinariamente dettagliate di caratteristiche geomorfologiche di varie zone del globo. In precedenza un altro sistema aveva volato sul veicolo spaziale *Seasat* nel 1978. Entrambe le apparecchiature sono state sviluppate dal Jet Propulsion Laboratory del California Institute of Technology.

In un sistema per immagini ottico o all'infrarosso l'energia emessa o riflessa da una superficie viene raccolta in un'apertura e messa a fuoco su un elemento (o su una schiera di elementi) che la rileva. Il potere risolutivo angolare del sensore è determinato dal rapporto fra la lunghezza d'onda di osservazione e il diametro dell'apertura; il potere risolutivo spaziale è pari al prodotto del potere risolutivo angolare per la distanza fra il sensore e la superficie. Il potere risolutivo diminuisce aumentando la quota del sensore o la lunghezza d'onda impiegata, oppure diminuendo il diametro dell'apertura. Poiché la lunghezza d'onda delle radiazioni visibili e infrarosse è molto corta è possibile ottenere immagini ad alta risoluzione anche da sistemi orbitanti ad altezze di diverse centinaia di chilometri con un'apertura di ragionevoli dimensioni.

Nella regione delle microonde la lunghezza d'onda è di diversi ordini di grandezza maggiore che non per le radiazioni visibili; entrambi i radar *SIR-A* e *Seasat* lavoravano su lunghezze d'onda di circa 24 centimetri, grosso modo un milione di volte maggiori della lunghezza d'onda della luce visibile. Occorre moltiplicare dello stesso fattore l'apertura, perché un radar installato su satellite raggiunga lo stesso potere risolutivo spaziale di un sistema ottico posto alla stessa quota. Per esempio, per ottenere una risoluzione sul terreno di 25 metri un radar per immagini convenzionale, a 250 chilometri di altezza sopra la superficie terrestre e funzionante a una lunghezza d'onda di 20 centimetri, avrebbe bisogno di un'apertura di ben due chilometri! È chiaro quindi che occorre sfruttare qualche altra tecnica per ottenere immagini ad alta risoluzione con un sensore radar posto a grandi distanze.

Una rappresentazione di un complesso uso del territorio è riportata in questa immagine radar a falsi colori di una zona rurale della provincia di Hopeh nella Cina nordorientale. L'immagine è stata registrata dal sistema *SIR-A (Shuttle Imaging Radar-A)*, da un'altezza orbitale di circa 250 chilometri. L'elaborazione del colore è stata eseguita al Jet Propulsion Laboratory del Cal Tech. I colori non hanno alcuna relazione con i colori naturali della scena, ma sono stati attribuiti soltanto in funzione dell'intensità delle radiazioni a microonde riflesse rilevata dall'antenna radar. Le macchie rosse corrispondono a villaggi; le aree in verde scuro sono campi coltivati e le linee in verde chiaro strade e canali di irrigazione; le zone in azzurro sono laghi e stagni, alcuni dei quali formati da dighe poste lungo i corsi d'acqua. La coltivazione prevalente nella zona è il grano.

In questa immagine a falsi colori del *SIR-A*, relativa a una regione arida e montuosa nella provincia del Sinkiang, predominano caratteristiche geologiche a grande scala. Le catene montuose sono state generate dal sollevamento di rocce sedimentarie; gli spostamenti lungo le catene indicano le principali zone di faglia. I conoidi alluvionali formatisi per processi erosivi lungo i fianchi delle montagne sono molto scabrosi alla lunghezza d'onda del radar (24 centimetri) e rinviano forti segnali riflessi. Le *playa* (bacini desertici) fra le catene sono costituite da un materiale più fine e regolare che dà luogo a echi radar relativamente deboli. Nelle zone montuose sono visibili pieghe e stratificazioni delle rocce.

In questa immagine radar di una regione coperta da fitte foreste nel Messico meridionale in prossimità del confine con il Guatemala compaiono caratteristiche geologiche fini. L'immagine è stata registrata dal sistema radar a bordo del satellite *Seasat* nel 1978. I dati sono stati elaborati con speciali tecniche per modulare nell'immagine il colore in funzione delle variazioni di bassa frequenza (ovvero di grande scala) e la luminosità in funzione di quelle di alta frequenza (di piccola scala). È così possibile evidenziare caratteristiche non appariscenti, come ad esempio le faglie. La confluenza dei fiumi Lacantum *(in alto)* e Salinas *(in basso)* è appena fuori dell'immagine verso nord-ovest *(a destra)*.

La soluzione, che è stata sviluppata originariamente per i sistemi radar di bordo a visione laterale, è denominata radar ad apertura sintetizzata: un'antenna radar di dimensioni relativamente ridotte viene fatta funzionare come un'antenna assai grande (sebbene lunga e sottile) spostando il movimento dell'antenna lungo una traiettoria ben definita. Un'antenna di grandi dimensioni reali (oppure una schiera di antenne) trasmette segnali e riceve i loro echi contemporaneamente su tutta la sua superficie. Gli echi raccolti su ogni parte dell'antenna vengono combinati fra loro, una volta giunti nella rete in guida d'onda del radar, prima di essere rilevati dal ricevitore. La ragione per cui il potere risolutivo migliora con l'aumento dell'apertura non è dovuta solo al fatto che un'antenna di maggiori dimensioni può raccogliere un segnale più forte. È invece importante la caratteristica che ogni punto della superficie illuminata disperde l'eco verso le varie zone dell'antenna; i segnali di ritorno, quando si combinano, interferiscono reciprocamente secondo le ampiezze e le fasi relative. Le conseguenti figure di interferenza contengono in codice informazioni dettagliate della superficie.

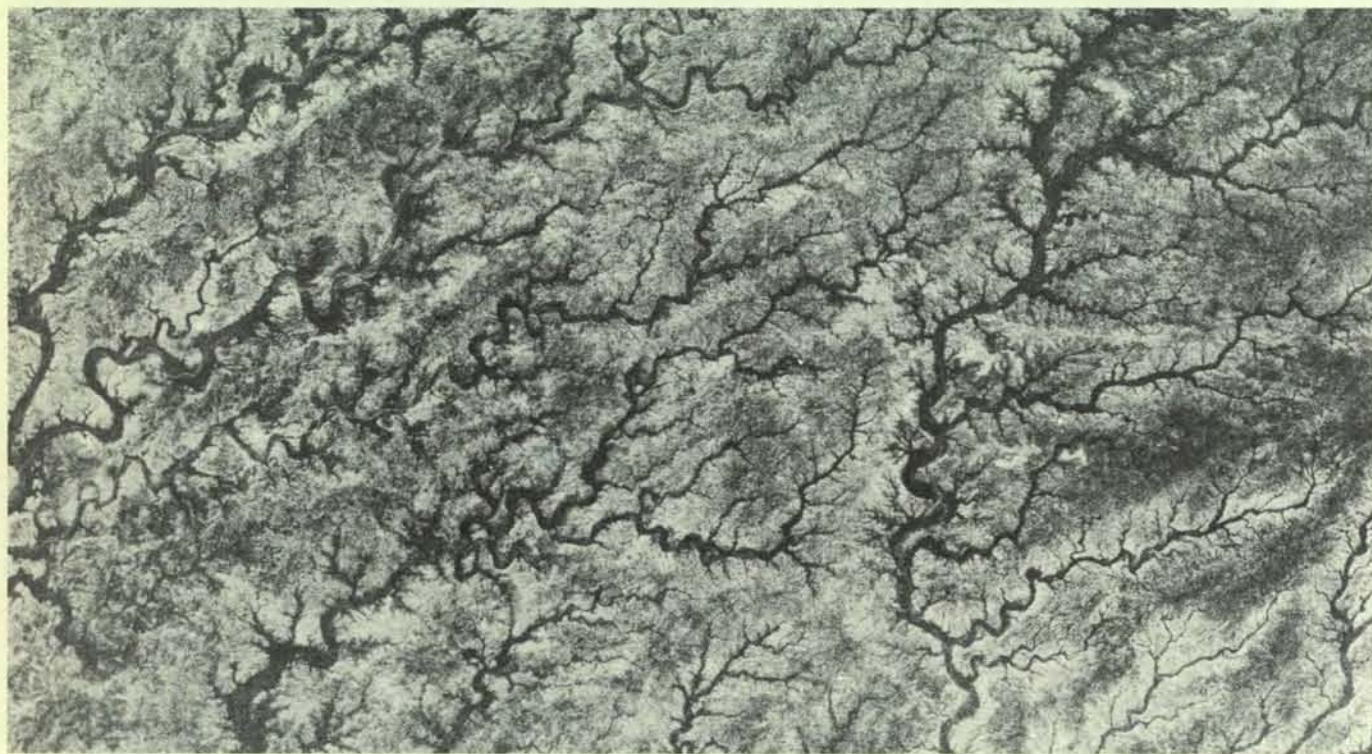
Anche un radar ad apertura sintetizzata risolve dettagli molto fini perché le informazioni sono ricavate da onde ricevute su un'area ampia. La differenza sta nel fatto che le onde non sono rilevate simultaneamente su tutta la superficie dell'antenna «sintetizzata»: i segnali ricevuti in ciascuna posizione occupata durante l'avanzamento lungo la traiettoria prestabilita vengono registrati e solo in seguito collazionati da un sistema di elaborazione dati. Per stabilire un riferimento mediante il quale sia possibile misurare l'ampiezza e la fase degli echi, si sovrappone un segnale di alta stabilità agli echi ricevuti dal sensore. In radioastronomia si usa una tecnica analoga per realizzare un interferometro a lunghissima base, mediante la combinazione di segnali ottenuti da una schiera di radiotelescopi posti a grande distanza l'uno dall'altro.

Il radar per immagini a bordo di satelliti spaziali funziona in base al principio dell'apertura sintetizzata, come viene illustrato qui. Il sistema radar a visione laterale emette un fascio di impulsi coerenti a microonde in direzione obliqua verso la superficie della Terra, mentre il veicolo si sposta lungo la sua rotta. I segnali riflessi ricevuti dall'antenna del radar vengono quindi registrati. Un sistema di elaborazione dei dati combina poi i segnali rilevati in corrispondenza dei vari punti della rotta orbitale per formare un'immagine bidimensionale dell'area interessata. L'antenna radar in movimento funziona come una singola antenna di grande lunghezza (oppure come una lunga schiera di piccole antenne); la massima lunghezza dell'apertura sintetizzata risultante è la distanza lungo la rotta in corrispondenza della quale un dato punto si trova all'interno del fascio emesso dall'antenna. Ogni punto della superficie è caratterizzato dalla distanza, che è funzione del ritardo fra l'istante di trasmissione dell'impulso e quello di ricezione dell'eco, e dall'azimut, determinato dalla misura della variazione di frequenza provocata dal movimento relativo fra veicolo spaziale e oggetto (effetto Doppler). L'immagine viene costruita mediante l'analisi di tutte le informazioni di ritardo e di spostamento Doppler codificate negli echi raccolti.

In pratica avviene che una piccola antenna, installata a bordo di un aeroplano o di un veicolo spaziale, si sposta secondo la rotta del supporto: la lunga schiera di antenne equivalenti è sintetizzata mediante la combinazione degli echi ricevuti nelle varie posizioni lungo la rotta. La massima lunghezza dell'apertura sintetizzata corrisponde alla distanza, misurata lungo la rotta, coperta durante il tempo in cui un determinato punto rimane all'interno del raggio «illuminante» emesso dall'antenna. La massima lunghezza dell'apertura sintetizzata dell'antenna aumenta proporzionalmente alla quota, di modo che il potere risolutivo spaziale ottenibile rimane costante (si veda l'articolo *Radar a visione laterale per il rilevamento topografico* di Homer Jensen, L. C. Graham, Leonard J. Porcello ed Emmett N. Leith in «Le Scienze» n. 113, gennaio 1978).

Il funzionamento di un qualsiasi sistema radar ad apertura sintetizzata è basato sul fatto che le proprietà delle onde che costituiscono il fascio illuminante possono essere stabilite a priori e controllate, cosa ovviamente impossibile per la luce solare. L'illuminazione è costituita da una serie di impulsi coerenti emessi su una frequenza a microonde, costante e stabile. Il tempo intercorrente fra la trasmissione dell'impulso e la ricezione della relativa eco fornisce la distanza del punto illuminato. I segnali che provengono da punti della superficie che si trovano a distanze diverse possono essere quindi identificati in base al tempo di arrivo dei loro echi sull'antenna. Tuttavia, per ciascuna data posizione del veicolo, esistono molti punti che si trovano alla medesima distanza e che formano un cerchio il cui centro è il piede della verticale dal satellite. È possibile distinguere fra loro i segnali provenienti da punti equidistanti misurando lo spostamento di frequenza degli echi per effetto Doppler, cioè la variazione di frequenza prodotta dal movimento relativo fra veicolo e terreno. Allo spostamento Doppler contribuisce solo la componente del moto parallela alla direzione del fascio a microonde e quindi l'eco proveniente da un punto sulla perpendicolare alla traiettoria di volo ha uno spostamento Doppler nullo. I punti più avanzati, nella direzione di volo, rispetto alla perpendicolare danno echi spostati verso frequenze superiori e quelli più arretrati danno echi spostati verso frequenze inferiori. I dati relativi al ritardo e allo spostamento Doppler degli echi, registrati durante il sorvolo di una determinata area, sono elaborati per isolare l'energia riflessa da ogni elemento risolto sulla superficie. Pertanto l'analisi delle informazioni di ritardo e di spostamento Doppler associate agli echi consente di costruire una mappa dell'energia ricevuta sotto forma di un'immagine bidimensionale che rappresenta punto per punto l'energia riflessa verso l'antenna.

Il potere risolutivo di un radar ad apertura sintetizzata è determinato dalla capacità del sensore di misurare con esat-

tezza minime differenze di ritardo e di spostamento Doppler rilevabili negli echi provenienti da elementi superficiali adiacenti. La precisione delle misurazioni dipende dalla distanza fra sensore e superficie e quindi, fintantoché il livello del segnale di ritorno si mantiene sufficientemente al di sopra del livello di rumore, il potere risolutivo, raggiungibile con tale tipo di sistema per immagini, è sostanzialmente indipendente dall'altezza. Entrambi i radar *SIR-A* e *Seasat* hanno fornito immagini con un potere risolutivo al suolo di 25 metri circa, impiegando un'antenna lunga 10 metri su una lunghezza d'onda di 24 centimetri. È un potere risolutivo che supera di un fattore tre quello raggiunto con i primi sensori Landsat, i quali sfruttavano la radiazione solare riflessa nel visibile e nell'infrarosso vicino.

I canali di drenaggio in un deserto roccioso lungo il confine fra Iraq e Arabia Saudita sono a malapena visibili in un'immagine Landsat *(in alto)*, ma sono nettamente definiti in un'immagine radar *SIR-A* *(in basso)*. Corsi d'acqua intermittenti tagliano le strette valli chiamate uadi. Gli uadi sono parzialmente riempiti di sabbia portata dal vento, che riflette un segnale molto più debole verso il radar che non la roccia circostante. Il forte contrasto nell'immagine radar facilita l'identificazione dell'andamento dei canali rispetto a quanto si verifica nell'immagine Landsat.

L'analisi del ritardo e dello spostamento Doppler, associati agli echi registrati mentre un singolo elemento è visto dal sensore, richiede un gran numero di calcoli e la manipolazione di molti dati. Nel caso del radar *Seasat* occorrevano qualcosa come 1000 operazioni complesse per sintetizzare la lunga apertura e per determinare la luminosità di un singolo pixel (elemento d'immagine) di 25 metri di lato. Il satellite si spostava alla velocità di circa 7,5 chilometri al secondo e il sensore copriva una striscia larga al suolo 100 chilometri: occorrevano quindi circa un miliardo di operazioni per generare un'immagine ad alta risoluzione dalle informazioni raccolte in un secondo. In più l'elaboratore dei dati doveva tener conto sia degli spostamenti Doppler provocati dalla rotazione della Terra sia delle distorsioni introdotte da leggere variazioni di quota e di assetto del sensore.

La necessità di grosse capacità di elaborazione per i radar ad apertura sintetizzata installati a bordo di satelliti è stata uno dei principali ostacoli allo sviluppo di questi sistemi. Il CRAY-1, uno dei più veloci calcolatori oggi disponibili, è in grado di eseguire circa 100 milioni di operazioni al secondo, velocità di calcolo che è di un ordine di grandezza inferiore a quella occorrente per elaborare in tempo reale i dati provenienti dai radar *SIR-A* o *Seasat*. Sono stati sviluppati due metodi per affrontare l'elaborazione di questa enorme quantità di dati: un metodo ottico e uno digitale.

La registrazione ottica degli echi di ritorno da un oggetto puntiforme su un'emulsione fotografica mostra una serie caratteristica di anelli concentrici simili alle figure di diffrazione di Fresnel che si incontrano abitualmente in ottica. Un'onda ottica piana incidente su tale diagramma può essere messa a fuoco ottenendo un'immagine del soggetto. Questo semplice procedimento (che può essere pensato come un analogo bidimensionale dell'olografia) è alla base di tutte le tecniche ottiche per l'elaborazione di dati provenienti da un radar ad apertura sintetizzata. Gli elaboratori ottici funzionanti su questo principio furono sviluppati originariamente fra gli anni cinquanta e l'inizio degli anni sessanta da ricercatori dell'Università del Michigan e della Goodyear Aerospace Corporation per essere applicati all'elaborazione di dati provenienti da radar ad apertura sintetizzata installati a bordo di aerei. I sistemi ottici furono in seguito adattati dal Jet Propulsion Laboratory all'elaborazione dei dati ricavati dai radar *SIR-A* e *Seasat*.

Il secondo metodo di elaborazione fa assegnamento esclusivo sui calcolatori digitali. L'attuale capacità è limitata, a causa dell'enorme numero di calcoli, alle operazioni che non cercano di mantenersi al passo con il flusso di dati proveniente dal sensore. Il più veloce degli elaboratori digitali disponibili per i radar per immagini di bordo ha un «rapporto di esecuzione» di uno a 500, intendendosi con questa dizione che i dati raccolti in un secondo richiedono 500 secondi di elaborazione. L'elaboratore è stato sviluppato presso il Jet Propulsion Laboratory da un gruppo guidato da Chialin Wu. Altri elaboratori meno veloci sono stati realizzati in Gran Bretagna, Canada, Germania e Giappone. Un elaboratore in grado di elaborare i dati provenienti dal sensore con un rapporto di esecuzione di uno a uno è in fase di sviluppo presso il Jet Propulsion Laboratory: il completamento è previsto per il 1985.

Le variazioni di struttura sono state analizzate matematicamente per codificare in falsi colori un'immagine del *SIR-A* di una regione montuosa del Belize. Differenze strutturali nelle varie parti dell'immagine corrispondono a variazioni spaziali riscontrate nella conformazione del terreno. Thomas G. Farr, che fa parte del gruppo di lavoro dell'autore dell'articolo, sta studiando questa tecnica quale modo di rappresentazione delle aree geografiche secondo i tipi di roccia.

Le immagini radar della Terra riprese dallo spazio trovano applicazioni in varie discipline: in geologia, in oceanografia, in planetologia e nello studio delle risorse rinnovabili. Un'applicazione in cui lo sfruttamento dei dati radar è particolarmente avanzato è la rappresentazione cartografica degli aspetti del terreno. Molte caratteristiche geologiche in superficie o in prossimità della superficie, quali faglie, duomi, affioramenti rocciosi e altre formazioni analoghe, sono associate con variazioni identificabili nell'aspetto della superficie e in modo specifico nella sua topografia, nella sua scabrosità o nella copertura vegetale. Anche i processi di erosione determinano caratteristiche e formazioni identificabili. Poiché la riflessione delle microonde è molto sensibile ai cambiamenti nelle proprietà fisiche della superficie, le caratteristiche e le formazioni sono chiaramente visibili e in alcuni casi intensificate nelle immagini radar.

Per ogni lunghezza d'onda l'osservazione dei dettagli topografici è dipendente in modo molto marcato dalla geometria dell'illuminazione. La maggiore facilità di osservazione dei dettagli si ottiene con illuminazione pressoché perpendicolare alla direzione dell'andamento topografico. Nel caso di sensori nel visibile o nell'infrarosso vicino la direzione di illuminazione è determinata dalla posizione del Sole e varia soprattutto in direzione est-ovest. Nel caso dei sensori radar la direzione di illuminazione e l'angolo di incidenza possono essere invece scelti e controllati, il che assicura una maggiore flessibilità. Un cambiamento di pochi gradi dell'angolo di incidenza può variare l'energia riflessa di un fattore due o più: i sensori radar sono perciò lo strumento ideale per rilevare minute caratteristiche strutturali che influenzano la topografia. Questa proprietà delle immagini radar è particolarmente utile nello studio di regioni coperte da fitte foreste, come quelle che si trovano ai tropici. Sono state sviluppate particolari tecniche di elaborazione delle immagini radar per mettere in evidenza caratteristiche topografiche fini allo scopo di esaltare sia le variazioni spaziali di bassa frequenza (cioè di grande scala) sia quelle di alta frequenza (di piccola scala).

I processi di erosione agiscono su varia scala sulla geomorfologia. Le scabrosità superficiali comprese fra pochi centimetri

e alcuni metri dipendono dalla composizione del terreno e dai particolari processi erosivi e, pur avendo solo un effetto limitato sul fattore di riflessione ottico del terreno, sono elemento predominante nel determinare la quantità di energia radar riflessa. Questo effetto si è rivelato particolarmente utile nel tracciare mappe di regioni molto aride dove il deserto «maschera» la superficie con sottili strati di sabbia che tendono ad abbassare il potere riflettente e a ridurre le variazioni spettrali nelle regioni del visibile e dell'infrarosso vicino dello spettro.

A causa della notevole lunghezza d'onda della radiazione sfruttata dai sensori radar sarebbe logico attendersi una certa penetrazione al di sotto della superficie. La profondità di penetrazione è di solito proporzionale alla lunghezza d'onda del segnale ed è fortemente influenzata dall'umidità superficiale. Nelle regioni molto aride del nostro pianeta è in effetti possibile una penetrazione, al di sotto della superficie, di alcuni metri. Gli specialisti dell'US Geological Survey e del Jet Propulsion Laboratory, analizzando i dati acquisiti dal *SIR-A* relativi al Sahara orientale egiziano, hanno di recente osservato che molte caratteristiche di grande scala rilevate dalle immagini radar non trovano riscontro né nelle immagini Landsat né al suolo. Dovrebbe trattarsi quindi di caratteristiche che si trovano al di sotto della superficie, ricoperte da uno strato di sabbia dello spessore di qualche metro. I geologi impegnati nello studio del Sahara sospettavano da tempo la presenza di tali caratteristiche; esse comprendono valli fluviali (alcune larghe quanto l'attuale valle del Nilo), terrazze fluviali, bacini desertici e letti rocciosi. La capacità dei sensori radar di penetrare al di sotto della superficie, recentemente confermata mediante ricerche al suolo, promette di avere conseguenze di grande portata per l'esplorazione archeologica, geologica e idrologica di regioni aride.

I segnali a microonde possono penetrare per uno spessore limitato anche nel manto di vegetazione che ricopre le zone umide tropicali. Immagini del *SIR-A* relativo al Borneo meridionale sembrano delineare aree paludose completamente ricoperte da vegetazione.

Questa immagine radar simulata del Monte Shasta nella California settentrionale è stata preparata da Michael Kobrick del Jet Propulsion Laboratory nel corso di una ricerca volta a valutare la possibilità di desumere informazioni tridimensionali sulla superficie terrestre mediante un interferometro radar montato su satellite. In un sistema del genere la differenza di fase fra gli echi a microonde ricevuti da due antenne montate sullo stesso veicolo spaziale verrebbe misurata punto per punto al fine di generare un'immagine topografica. In questo caso i dati topografici sono stati in effetti ottenuti prima con mezzi convenzionali e quindi sono stati elaborati per simulare l'uscita digitale dell'interferometro radar orbitante. Le fasce di diverso colore, derivate dal flusso risultante di informazioni digitali, corrispondono alle isoipse di una cartina topografica.

Risultati analoghi sono stati ottenuti presso l'Università dell'Arkansas dall'analisi di immagini radar del *Seasat* relative all'Arkansas centrale.

La discriminazione e l'identificazione di differenti tipi di rocce mediante telerilevamento può essere ottenuta nel migliore dei modi analizzando la «firma» spettrale della radiazione visibile e infrarossa riflessa o emessa. Recenti sviluppi dovrebbero consentire l'estensione di tali rappresentazioni spettroscopiche bidimensionali a un gran numero di bande spettrali nel visibile e nell'infrarosso per diagnosticare la composizione delle rocce superficiali. In questo campo le immagini radar possono fornire informazioni complementari offrendo un contributo indiretto alla stesura di mappe litologiche. La scabrosità superficiale e le strutture topografiche di piccola scala, che hanno importanti effetti sull'eco radar, dipendono dal tipo di rocce presenti: l'informazione radar è quindi utile per discriminare, anche se non per identificare, i tipi di roccia. Così, per esempio, quando una zona litologica è stata individuata con altri mezzi, la stesura di una mappa radar può aiutare a definirne i confini.

Il lavoro compiuto dal mio gruppo al Jet Propulsion Laboratory ha dimostrato che l'aggiunta dei dati relativi alle immagini radar a quelli forniti dall'apparecchiatura a scansione multispettrale del Landsat migliora la capacità di classificazione delle rocce. L'importanza dei dati radar è perfino maggiore in zone ricoperte da vegetazione ove non può essere ottenuta un'immagine multispettrale. In queste aree l'unica via praticabile per classificare a distanza le rocce consiste nello sfruttare le variazioni della struttura superficiale e della capacità di drenaggio, due proprietà che sono in genere connesse con il tipo di roccia. Presso il Jet Propulsion Laboratory e presso l'Università del Kansas sono in corso di sperimentazione alcune tecniche di elaborazione per giungere a classificare le aree in base alla struttura osservata nelle immagini radar. Gli sforzi sono stati particolarmente fruttuosi nello studio di regioni ricoperte da fitta vegetazione.

La presenza di vegetazione o di strutture costruite dall'uomo ha un considerevole effetto sugli echi radar. La vegetazione tende a disperdere fortemente le onde radar a causa sia dell'elevato contenuto di umidità sia del grandissimo numero di superfici riflettenti. I campi coltivati e le radure artificiali nei boschi vengono delineati con chiarezza nelle immagini radar. Sono in corso studi presso l'Università del Kansas sul problema di correlare l'intensità del segnale radar riflesso con la natura, lo stadio di crescita e le condizioni di «salute» della vegetazione. I tentativi di correlare le misurazioni al suolo con le immagini radar hanno dato finora risultati incoraggianti.

Le strutture costruite dall'uomo riflettono fortemente i segnali radar in primo luogo perché le superfici piane perpendicolari tendono a formare dei riflettori angolari che restituiscono la maggior parte dell'energia incidente e poi perché le strutture metalliche si comportano come antenne che reirradiano fortemente l'energia incidente. Nelle immagini del *Seasat* e del *SIR-A* sono visibili molte strutture artificiali, anche quando hanno dimensioni inferiori al potere risolutivo. Le caratteristiche osservate più di frequente comprendono i tracciati ferroviari, le strade, i tralicci degli elettrodotti, i ponti, le piattaforme petrolifere e i natanti. Si sta mettendo alla prova la possibilità di combinare immagini radar e immagini Landsat per migliorare la capacità di controllare a distanza su larga scala l'uso urbano del territorio.

Tutti i sensori per immagini, comprese le apparecchiature radar, forniscono una rappresentazione bidimensionale: la topografia di una superficie non può quindi essere completamente desunta da una singola immagine. Per determinare l'altezza dei vari elementi e per ottenere una carta topografica con curve di livello è necessaria una coppia stereoscopica di immagini riprese da posizioni differenti. Il procedimento per la costruzione di mappe topografiche partendo da coppie di immagini stereoscopiche potrebbe essere impiegato con le immagini radar, e in effetti sono in corso studi in questa direzione. Il metodo è però tedioso, in particolare per la realizzazione di mappe a grande scala.

In questa immagine radar *Seasat* delle acque intorno all'isola Kayak nell'Alaska meridionale, sono evidenti le onde lunghe oceaniche, con lunghezza d'onda di circa 200 metri. Le onde sono rifratte e diffratte quando interagiscono con la costa dell'isola e il basso fondale della Controller Bay.

Il fatto che il sensore radar fornisca e controlli l'illuminazione ha indotto L. C. Graham della Goodyear Aerospace a introdurre una tecnica interferometrica da impiegare in unione con quella dell'apertura sintetizzata per ottenere l'informazione tridimensionale della superficie. La soluzione proposta prevede l'installazione sullo stesso velivolo di due antenne separate. Delle antenne una sola illumina la superficie, ma entrambe raccolgono il segnale riflesso. I due ricevitori impiegano la stessa sorgente di riferimento per confrontare le fasi delle coppie di echi ricevuti. L'informazione sulla fase consente di misurare direttamente l'altezza di ciascun pixel ed elimina la necessità di un'analisi stereoscopica. Studi recenti e una dettagliata simulazione su modello presso il Jet Propulsion Laboratory indicano che la soluzione proposta per questo problema è applicabile anche alle riprese dallo spazio.

I segnali a microonde non penetrano nell'acqua per spessori significativi. D'altra parte l'intensità della riflessione radar è sensibile alle onde capillari superficiali e alle piccole onde di gravità, tanto quanto alla scabrosità superficiale del terreno. Pertanto in linea di principio qualsiasi aspetto degli oceani che modifichi lo stato della superficie può essere rilevato dai radar su satellite. Molte caratteristiche influenzano lo stato della superficie dei corsi d'acqua. Le onde lunghe provocano una modulazione periodica dell'intensità delle onde più piccole e così nelle immagini radar si può osservare comunemente l'andamento delle onde lunghe. Anche la velocità delle correnti modula lo stato della superficie. A sua volta la velocità delle correnti è modulata dalla topografia del fondale, là dove le acque sono basse; per questo in numerose immagini radar sono state rilevate le caratteristiche del fondale in prossimità delle coste, benché in realtà i segnali a microonde non raggiungano mai il fondale stesso.

Le onde interne si propagano nell'oceano lungo disomogeneità stratificate. La circolazione dell'acqua connessa con tali onde modula le irregolarità superficiali secondo configurazioni spaziali ben definite e riconoscibili. Le immagini eseguite dal *Seasat* e dal *SIR-A* hanno mostrato che le onde interne sono molto più comuni di quanto si pensasse. Lo stato della superficie è influenzato anche da vortici, anelli di acqua calda, fronti, sedimenti in sospensione e celle di vento in prossimità della superficie. Questi fenomeni, osservabili tutti nelle immagini radar, possono essere seguiti indipendentemente dalle condizioni meteorologiche, con cielo sereno o coperto, di giorno e di notte. I radar a bordo di satelliti possono quindi fornire coperture ripetitive per studiare le variazioni delle caratteristiche superficiali e per seguire i loro movimenti.

Oltre all'effetto dell'irregolarità superficiale sull'intensità del riflesso radar dall'oceano, il movimento superficiale influenza la frequenza del segnale riflesso stesso, per effetto Doppler. Ricercatori del Jet Propulsion Laboratory e dell'Environmental Research Institute of Michigan stanno studiando lo sfruttamento di questo piccolo spostamento Doppler per rappresentare in mappa la velocità delle correnti oceaniche.

I banchi di ghiaccio galleggianti, le creste ghiacciate e i canali liberi negli oceani polari forniscono riflessioni radar caratteristiche, ben distinguibili nelle immagini. Coperture ripetute eseguite nell'estate del 1978 con il radar *Seasat* ad apertura sintetizzata sul Mare di Beaufort hanno consentito a Benjamin Holt del Jet Propulsion Laboratory di seguire gli spostamenti di una formazione di ghiaccio per oltre 160 chilometri. Un metodo del genere dovrebbe permettere di controllare a lungo termine i movimenti dei ghiacci polari trascinati dalle correnti che si pensa compiano un giro completo intorno al polo in due anni. Il controllo continuo dei banchi di ghiaccio galleggianti è inoltre necessario per pianificare le rotte marine e per decidere nel migliore dei modi la localizzazione delle piattaforme petrolifere nei mari polari. Questa possibilità ha indotto il governo canadese ad approntare un satellite per l'osservazione dei ghiacci dotato di un sistema radar ad apertura sintetizzata da lanciare alla fine degli anni ottanta. Anche la delimitazione della calotta polare e della sua estensione hanno un particolare interesse per determinare il flusso di calore dagli oceani all'atmosfe-

ra, una variabile essenziale nei modelli del clima mondiale.

Fino a poco tempo fa l'analisi delle immagini radar era soprattutto qualitativa e basata sulla identificazione di formazioni e aspetti particolari, sulla misura delle loro dimensioni e sullo studio delle associazioni di forme particolari, per desumerne informazioni sulla superficie. Negli ultimi anni è stato avviato lo sviluppo di tecniche quantitative che dovrebbero fornire una migliore comprensione della «firma» che lasciano, nelle immagini radar, le diverse caratteristiche superficiali. Immagini radar calibrate potrebbero fornire ulteriori informazioni sulla scabrosità e sull'umidità del suolo, mentre tecniche matematiche potrebbero applicarsi alla classificazione delle strutture superficiali. I dati quantitativi forniti da immagini radar e da immagini multispettrali potrebbero facilitare l'identificazione di varie caratteristiche. Una serie di immagini radar eseguite con fasci aventi piani di polarizzazione differenti potrebbe rendere possibile l'effettuazione di misure separate della scabrosità superficiale e della costante dielettrica (una proprietà elettrica fondamentale). Le stesse immagini riprese con frequenze differenti potrebbero essere impiegate anche per misurare la gamma delle scabrosità superficiali. Molte tecniche sviluppate in origine per l'interpretazione delle immagini Landsat vengono ora applicate alle immagini radar, consentendo così di ottenere rapidi progressi nello sviluppo di metodi radar.

Sono in corso vari programmi per sviluppare sistemi radar di osservazione più raffinati. Negli Stati Uniti si pone l'accento su sensori radar polivalenti a fini di ricerca, che saranno portati in orbita dallo Space Shuttle. Il programma statunitense prevede quale prossima fase il volo del *SIR-B* nel 1984, da cui si otterranno immagini con numerose differenti geometrie di illuminazione. In Canada, Europa e Giappone l'attenzione è maggiormente rivolta allo sviluppo di sistemi in orbita per osservazioni a lungo termine.

Il sensore radar è l'unico strumento disponibile per osservare la superficie di alcuni corpi celesti nel sistema solare. Sappiamo ben poco della superficie di Venere e di Titano (il maggiore satellite di Saturno), a causa della continua e totale copertura di nubi. Durante la recente missione Pioneer Venus Orbiter sono state ottenute immagini radar con una risoluzione di circa 80 chilometri che hanno fornito il primo sguardo su molte delle principali caratteristiche superficiali del pianeta. Una missione Venus Radar Mapper pianificata, ma non ancora approvata, potrebbe un giorno fornire un quadro molto più dettagliato della sua superficie. Nel caso di Titano le conoscenze della sua superficie sono nulle e oggetto di molte ipotesi. Un radar per immagini in orbita intorno al satellite è la sola scelta disponibile per costruire un quadro completo della superficie di questo remoto mondo.

Lo sviluppo nel cervello di mappe e strie

Nel cervello umano le cellule nervose formano, per i loro rapporti con il mondo esterno, delle mappe che sono divise in strie. Queste ultime possono essere analizzate ricorrendo a una rana con tre occhi

di Martha Constantine-Paton e Margaret I. Law

Il cervello di un vertebrato è la struttura più complessa presente in un qualsiasi organismo vivente. La versatilità e le capacità analitiche di quella struttura sono suggerite, al microscopio, dall'aspetto dei singoli neuroni. Ognuna di queste cellule appare diversa dalle cellule confinanti poiché presenta una forma elaborata e stabilisce legami con altri neuroni mediante sinapsi all'estremità dell'assone, o fibra nervosa. In questa diversità vi è, tuttavia, una notevole coerenza. Via via che nuove conoscenze si accumulano sui tipi di connessioni in varie parti del cervello, cominciano ad affiorare principi organizzativi, il che fa sorgere la speranza che i tipi di interazioni neuroniche, che sono alla base della evidente complessità del cervello, risultino poco numerosi, così da essere controllabili.

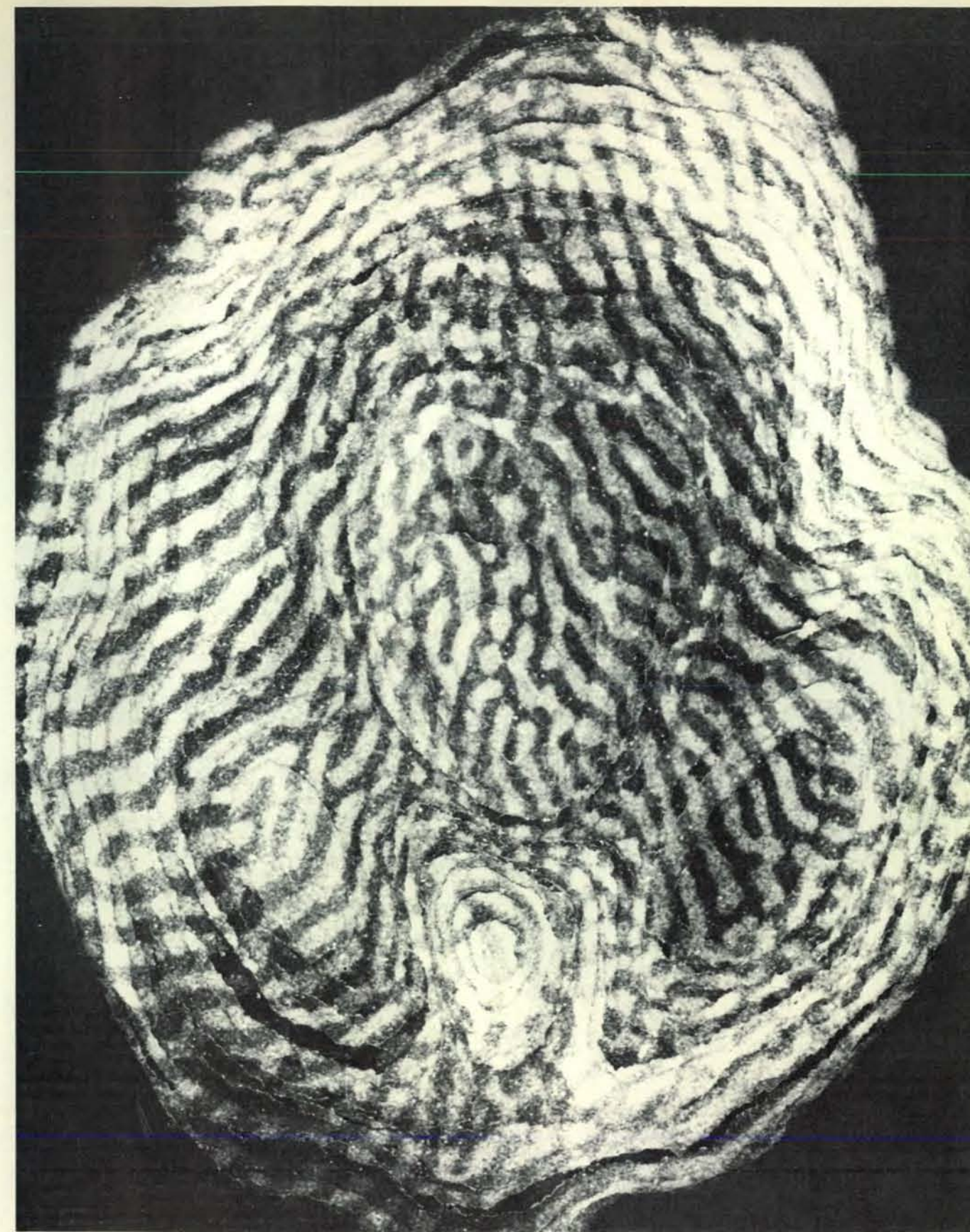
Uno di questi principi organizzativi è la formazione di mappe. Gli assoni che si dipartono da neuroni localizzati in una regione del cervello e che si proiettano verso i neuroni di un'altra regione in genere riproducono rapporti di vicinanza. Di conseguenza, se due neuroni sono vicini nella prima regione, le loro sinapsi si formeranno sulla stessa cellula, o su cellule vicine, nella seconda regione (popolazione bersaglio). Questa regolarità nelle proiezioni assoniche è stata individuata per la prima volta nel XIX secolo e oggi è stata trovata in tutte le proiezioni lungo le quali segnali sensoriali raggiungono la corteccia cerebrale, nelle proiezioni che servono a collegare un'area della corteccia cerebrale a un'altra e in quelle, infine, attraverso le quali le parti del cervello che sono interessate nel controllo del movimento agiscono sui muscoli corporei.

Un secondo principio organizzativo, scoperto molto più di recente, è la compartimentazione delle regioni del cervello che incorporano una mappa in sottodistretti periodici. Per esempio, ricerche condotte da Jon H. Kaas nel suo laboratorio alla Vanderbilt University hanno rivelato una compartimentazione estremamente regolare della corteccia sensoriale somatica, cioè di quella parte della corteccia cerebrale che riceve informazioni sensoriali dai muscoli, dalle articolazioni e dalla cute. In essa la superficie della mano è rappresentata da una mappa. Pertanto, se si toccano due punti sulla cute, che sono vicini l'uno all'altro, viene stimolata una attività elettrica misurabile in gruppi di neuroni vicini nella corteccia. In esperimenti effettuati su scimmie, Kaas e collaboratori hanno trovato che nello strato corticale in cui gli assoni che penetrano formano delle sinapsi, cioè nel cosiddetto «Strato 4», la mappa viene suddivisa in strie. Queste separano gli impulsi sensoriali provenienti dalla mano secondo il tipo di informazione che portano. In alcune strie i neuroni rispondono solo all'inizio del contatto; in quelle che si frappongono i neuroni danno una risposta più prolungata. È come se la mappa della mano nella corteccia sensoriale somatica della scimmia fosse stata costruita alternando strie tagliate da due distinte mappe della mano. Una mappa (e una serie di strie) rappresenta le terminazioni nervose della cute, caratterizzate da un rapido adattamento; l'altra mappa terminazioni nervose con un adattamento più lento.

Edward G. Jones e collaboratori, alla Washington University School of Medicine, sono stati tra i primi a dimostrare, sotto l'aspetto anatomico, che quelle suddivisioni funzionali hanno origine in quanto ogni suddivisione riceve assoni particolari. I ricercatori sopracitati si sono basati su una tecnica radiografica: hanno iniettato nella corteccia sensoriale somatica su un lato del cervello di una scimmia una piccola quantità di amminoacidi marcati con trizio, l'isotopo radioattivo dell'idrogeno; questi amminoacidi sono stati assunti dai neuroni nella corteccia e trasportati lungo gli assoni che si dipartono da alcuni di questi neuroni fino a giungere alla corteccia sensoriale somatica sul lato opposto del cervello. Qui gli assoni marcati si sono distribuiti secondo un certo schema, riconoscibile in base alla radioattività emanata da sezioni di tessuto cerebrale e evidenziata rivestendo le sezioni con un'emulsione fotografica sensibile a essa.

La distribuzione della radioattività nelle sezioni successive ha messo in evidenza che gli assoni terminavano in una serie chiaramente delimitata di strie. Così, i messaggi che passavano da un lato all'altro del cervello e quelli che portavano informazioni provenienti dalla cute andavano a finire in strie della corteccia sensoriale somatica. La ripartizione degli stimoli alla corteccia somatica sensoriale non è, però, l'unico esempio di distribuzione in strie. Vi sono, infatti, molti altri esempi: strie sono state trovate in tutte le vie sensoriali, in molte regioni della corteccia cerebrale e in regioni del cervello molto diverse tra loro, come il collicolo superiore, il cervelletto e il midollo allungato.

Le zone sinaptiche periodiche, che formano strie funzionali in una regione del cervello che incorpora simultaneamente una mappa, costituiscono un enigma. Perché si trovano lì? Perché il cervello dovrebbe mettere in atto un elaborato meccanismo di segregazione dei vari messaggi, quando in ultima analisi questi messaggi convergeranno così da produrre una rappresentazione unificata? Noi abbiamo compiuto una serie di esperimenti che suggeriscono una risposta a questo interrogativo.

Il nostro lavoro si concentra sul sistema visivo, una serie di proiezioni che portano l'informazione visiva dalla retina a stazioni più centrali del cervello. Queste vie sono state studiate in modo approfondito e, di conseguenza, si conosce meglio la rappresentazione del mondo visibile, fatta dal sistema nervoso centrale, di quanto non si conoscano le rappresentazioni di qualunque altra modalità sensoriale. Buona parte delle ricerche sono state realizzate, sul gatto e sulla scimmia, da David H. Hubel, Torsten N. Wiesel e dai loro collaboratori della Harvard Medical

Le strie, nel cervello di un mammifero, si trovano nella corteccia visiva, cioè in quella parte della corteccia cerebrale che riceve messaggi dagli occhi. La parte di tessuto qui mostrata è circa un quarto dell'intera corteccia visiva di un lato del cervello di un macaco. In uno degli occhi di questo animale è stata iniettata una piccola quantità di un amminoacido (prolina), marcato con trizio, l'isotopo radioattivo dell'idrogeno, e dopo due settimane la radioattività si è trasferita lungo gli assoni dall'occhio al nucleo genicolato laterale del cervello e da qui alla corteccia visiva. Quando sezioni di tessuto della corteccia visiva vengono rivestite con una emulsione fotografica, la radioattività dà luogo a strie chiare: queste si alternano a strie scure, che rappresentano l'occhio non iniettato. Ogni stria ha una larghezza di circa 350 micrometri. L'immagine è stata fornita da Simon LeVay della Harvard Medical School: si tratta di un montaggio prodotto con successive sezioni di tessuto.

School. I loro risultati costituiscono un'analisi dettagliata del rapporto esistente tra organizzazione topografica e organizzazione funzionale. Nel gatto e nella scimmia, le vie visive convogliano le informazioni alla parte di corteccia cerebrale designata come corteccia visiva: qui la mappa è binoculare, cioè deriva dagli assoni che portano le informazioni da ciascuno dei due occhi dell'animale. Hubel e Wiesel hanno trovato, però, che, facendo passare un microelettrodo attraverso il tessuto cerebrale dello Strato 4 della corteccia visiva, esso registra l'attività elettrica dei neuroni in una sequenza alternata, estremamente regolare. Una serie iniziale di neuroni risponderebbe solo a lampi di luce che colpiscono l'occhio sinistro dell'animale. Seguirebbe, quindi, una serie di cellule che rispondono solo a stimoli provenienti dall'occhio destro, e quindi ancora un'altra serie che risponde a stimoli dell'occhio sinistro.

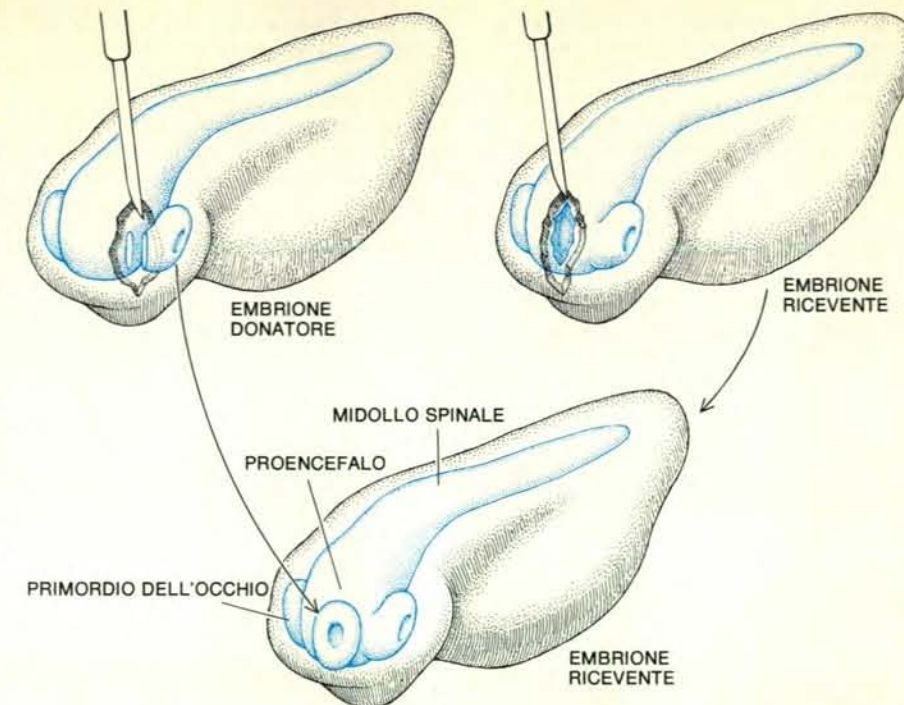
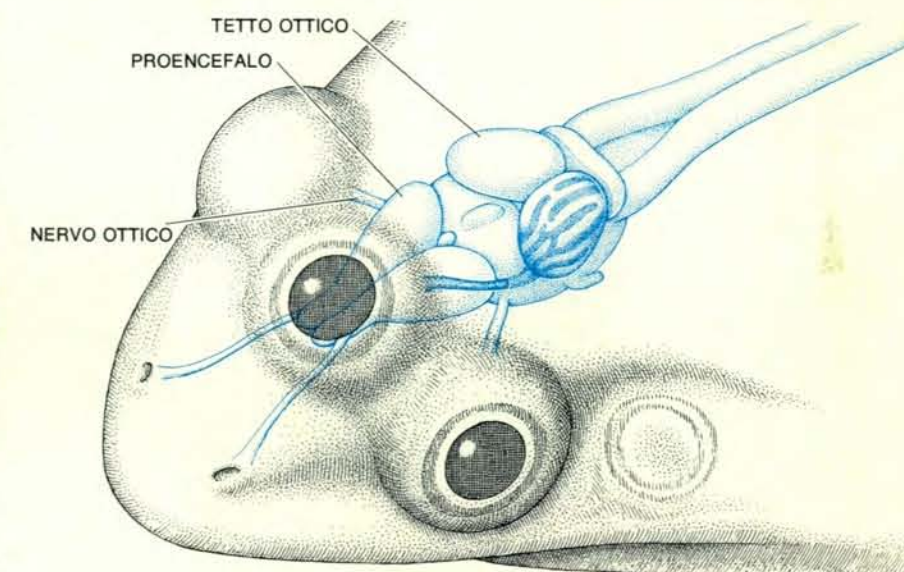
Hubel, Wiesel e collaboratori hanno, inoltre, dimostrato che tale alternanza funzionale deriva dalla segregazione degli assoni che convogliano le informazioni da ogni occhio. Marcando la via visiva di ogni occhio con amminoacidi radioattivi si rivela la presenza di strie che decorrono, attraverso lo Strato 4 della corteccia visiva, con un andamento a zebra. Ogni stria contiene cellule che reagiscono esclusivamente a un occhio e che, a loro volta, proiettano i loro assoni su neuroni binoculari, siti negli strati della corteccia sopra e sotto di loro. In altre parole, ogni parte del mondo visibile, su cui il gatto o la scimmia possono rivolgere ambedue gli occhi, è rappresentata due volte nello Strato 4: la prima volta in un certo punto all'interno di una stria, che rappresenta l'occhio sinistro, e la seconda in una stria vicina, che rappresenta l'occhio destro.

I nostri esperimenti realizzati alla Princeton University si sono avvalsi di parecchie proprietà di una via visiva notevolmente diversa, quella della rana leopardo *(Rana pipiens)*, basandosi su un procedimento classico di trapianto di tessuti in embrioni di anfibio. Questo trapianto, però, è stato combinato con tecniche moderne di neuroanatomia e neurofisiologia per poter esaminare i tipi di connessioni che i neuroni di un sistema visivo stabiliscono se posti in situazioni anomale all'inizio dello sviluppo. Talvolta queste analisi possono rivelare principi di crescita e di organizzazione che non risultano evidenti nel corso di uno sviluppo normale.

In una serie di esperimenti, abbiamo tolto un primordio di occhio (cioè il tessuto che diventerà un occhio) da embrioni, in uno stadio di sviluppo in cui l'occhio era rappresentato da una semplice estroflessione del sistema nervoso centrale. Abbiamo quindi trapiantato questo primordio in un secondo embrione, e precisamente nella regione dove si trovano i suoi due primordi. Gli embrioni così trattati sono diventati girini e quindi giovani rane con tre occhi, per il resto praticamente normali. L'occhio in più generalmente si trovava anteriormente a uno dei due occhi normali; talvolta si trovava, invece, sulla punta del naso o sulla sommità del capo.

Le rane leopardo dipendono molto dalla vista, ma diversamente dai gatti o dalle scimmie il loro cervello non ha sviluppato quella elaborata corteccia visiva che è caratteristica dei mammiferi. L'area principale in cui avviene l'elaborazione dell'informazione visiva si trova, invece, nel tetto ottico, una coppia simmetrica di lobi, che occupa buona parte del mesencefalo. Ogni lobo del tetto ottico riceve praticamente tutti gli assoni provenienti dalla retina di un occhio, l'occhio controlaterale (cioè che si trova sul lato opposto della testa). La proiezione da quest'occhio crea una mappa estremamente ordinata della superficie della retina, ma dato che il lobo non riceve una massiccia proiezione anche dalla seconda retina, non vi sono strie a rappresentare ogni occhio.

Le strie anomale nel cervello della rana leopardo *(Rana pipiens)* assomigliano in modo sorprendente a quelle che sono presenti nel cervello di un mammifero. Qui esse vengono messe in rilievo iniettando un enzima, la perossidasi di rafano, in uno dei nervi ottici dell'animale in modo che possa essere trasportato nel cervello dagli assoni che costituiscono quel nervo. Il cervello viene quindi trattato in modo che l'enzima produca, all'interno delle terminazioni degli assoni, una sostanza bruna. La fotografia in alto mostra una parte del cervello di una rana normale, vista da sopra. Ogni lobo è un lobo del tetto ottico, la regione cerebrale le cui cellule ricevono il nervo ottico dall'occhio posto sul lato opposto del capo. Il nervo ottico dell'occhio sinistro è stato iniettato con l'enzima; pertanto il lobo di destra è contrassegnato dal prodotto della reazione. Il colore uniforme suggerisce che gli assoni proiettati verso il tetto ottico distribuiscano in esso le loro terminazioni in modo continuo. Le cellule del tetto ottico includono così una mappa topografica della retina e, quindi, una mappa del mondo visibile. La fotografia in basso mostra il cervello di una rana anomala, una rana con tre occhi perché gli autori del presente articolo hanno trapiantato in essa un primordio soprannumerario di occhio (occhio prospettico), allo stadio di embrione. In questo soggetto, gli assoni dell'occhio soprannumerario sono stati iniettati con l'enzima; si possono vedere nel tetto ottico, dove le loro terminazioni si trovano in regioni che appaiono striate. Le strie si alternano con altre, dovute alle terminazioni di assoni provenienti dall'occhio normale. Ogni stria ha una larghezza di circa 200 micrometri.

Le rane con tre occhi sono diverse. Nella maggior parte di esse, la retina dell'occhio soprannumerario invia gli assoni prevalentemente a uno o all'altro dei due lobi del tetto ottico. Qui gli assoni soprannumerari competono con il messaggio normale che arriva al tetto, cioè con gli assoni che arrivano da uno degli occhi normali della rana. Abbiamo esaminato l'encefalo delle rane con tre occhi iniettando nell'uno o nell'altro dei due occhi normali o nell'occhio soprannumerario degli amminoacidi marcati con elementi radioattivi e abbiamo atteso un giorno o due perché l'isotopo avesse il tempo di essere trasportato, lungo gli assoni, fino alle loro terminazioni sinaptiche nel tetto. I lobi di quest'ultimo sono stati quindi sezionati e le singole sezioni sono state trattate per evidenziare la distribuzione delle terminazioni sinaptiche marcate.

Il procedimento chirurgico che produce una rana con tre occhi deve sottrarre il primordio di un occhio da un embrione per introdurlo in un secondo embrione dopo che da questo è stato rimosso del tessuto per fargli spazio. Al momento del trapianto ogni embrione ha una lunghezza di circa tre millimetri e ogni primordio di occhio è una estroflessione del proencefalo in via di sviluppo.

Una rana con tre occhi ha due occhi normali nella giusta posizione e il terzo posto o anteriormente a un occhio normale o sulla sommità del capo. In tre quarti dei casi, il terzo occhio compete con uno normale per la distribuzione delle terminazioni assoniche in uno dei due lobi del tetto.

Le due serie di terminazioni (marcate e no) non si mescolavano mai in una sezione. Al contrario, erano segregate in zone specifiche, che periodicamente si alternavano. Inoltre, seguendo le zone corrispondenti alle terminazioni marcate in sezioni successive si è visto che tali zone erano allineate in strie. Ogni stria aveva una larghezza di circa 200 micrometri e decorreva grosso modo dalla parte anteriore alla parte posteriore di ogni lobo, con un disegno a zebra, sempre uguale. Non contava se l'occhio soprannumerario proveniva originariamente dal lato destro o da quello sinistro di un embrione donatore, o se gli assoni di quest'occhio soprannumerario penetravano nel lobo destro o in quello sinistro del tetto ottico. La traiettoria seguita dagli assoni soprannumerari, il modo in cui si univano in fascio crescendo, e la direzione dalla quale entravano nel lobo del tetto ottico non facevano, essi pure, alcuna differenza.

In un esperimento correlato ai prece-

denti, parecchi laboratori, compreso il nostro, hanno rimosso uno dei due lobi del tetto ottico da una rana o da un carassio normale. Gli assoni che, dalla retina, si proiettavano sul lobo ora mancante si sono rigenerati e sviluppati nel lobo rimasto, dove hanno dovuto competere con le proiezioni che già vi si trovavano, producendo delle strie alternate di terminazioni derivate da assoni dell'occhio destro e dell'occhio sinistro. La duplicazione sperimentale dello stimolo visivo a una regione del tetto che, generalmente, sostiene solo una mappa della retina, sembra produrre inevitabilmente, in un vertebrato inferiore, una serie di suddivisioni funzionali complesse, sorprendentemente simili alla disposizione nella corteccia visiva dei mammiferi normali.

In alcune delle nostre rane con tre occhi abbiamo registrato l'attività in corrispondenza delle terminazioni degli assoni nel tetto ottico, mentre proiettavamo chiazze di luce su uno schermo di fronte alla rana. Abbiamo dapprima coperto gli occhi normali e quindi l'occhio soprannumerario. In questo modo, abbiamo potuto dimostrare che la rappresentazione di ciascun occhio sul tetto doppiamente innervato era allineata in modo giusto rispetto agli assi oculari originali dell'embrione. Così, gli assoni presenti in ogni proiezione conservavano nel tetto una mappa della retina da cui erano originati, anche se tale mappa risultava interrotta da strie alternantisi. Anche in questo caso le rane anomale assomigliano a un mammifero normale. Due stimoli retinici producono due rappresentazioni separate del mondo visivo in un'unica struttura bersaglio.

L'orientamento delle mappe nei lobi del tetto ottico di una rana con tre occhi viene determinato registrando l'attività elettrica di gruppi di assoni della retina, che terminano in vari punti di ogni lobo del tetto ottico, quando, su una superficie emisferica di fronte alla rana, sono proiettate delle luci. I numeri da 1 a 8 segnano i punti in cui, nel lobo del tetto ottico sul lato destro del cervello, sono state effettuate le registrazioni. Identificano anche i campi di ricezione: la parte del mondo visibile che le cellule incontrate in corrispondenza di ogni punto sono in grado di registrare. In questo caso, il lobo del tetto ottico sul lato destro del cervello include una mappa del mondo visto con l'occhio posto sul lato sinistro *(in grigio)*. Le frecce vuote, nella mappa, uniscono le varie direzioni del mondo visibile, rilevate dalle terminazioni assoniche disposte da parte a parte del lobo del tetto ottico; le frecce piene uniscono, invece, direzioni rilevate da terminazioni assoniche disposte dall'avanti all'indietro. Le lettere da *A* a *Y* segnano i punti nel lobo del tetto, sul lato sinistro del cervello, dove sono avvenute le registrazioni. Segnano anche i campi di ricezione. Il lobo sul lato sinistro risulta includere due mappe: quella dell'occhio normale *(in nero)* e quella dell'occhio soprannumerario *(in colore)*. Nella maggior parte dei punti di registrazione, le terminazioni assoniche che rappresentano ogni occhio sono abbastanza vicine da essere riconosciute mediante un singolo elettrodo. Ambedue le mappe sono ben organizzate per il fatto che, quando uno dei due occhi è coperto, l'attività delle cellule, in una successione di localizzazioni nel lobo del tetto, può essere sollecitata da stimoli provenienti da una successione di parti del mondo visibile. Ma la mappa normale e la mappa soprannumeraria non sono a registro: quando l'animale ha ambedue gli occhi aperti, gruppi vicini di cellule dei lobi del tetto ottico possono reagire a parti diverse del mondo visibile.

In una rana con tre occhi, in cui l'occhio soprannumerario invia assoni ad ambedue i lobi del tetto ottico, si trovano queste immagini, specularmente simmetriche, di strie. In particolare, un buco nella distribuzione delle strie in un lobo corrisponde a un insieme di strie, specularmente simmetrico, nell'altro. Le immagini suggeriscono che gli assoni provenienti dall'occhio soprannumerario competano, nel lobo del tetto ottico, con gli assoni provenienti da un occhio normale solo in corrispondenza di regioni particolari, determinate dal punto in cui gli assoni emergono dalla retina. La superficie di ogni lobo è stata ricostruita misurando la larghezza delle strie in una serie di sezioni del lobo del tetto, effettuate con intervalli di 20 micrometri fra l'una e l'altra.

In un mammifero, però, gli occhi hanno posizioni simmetriche sul capo. Non è questo il caso della rana con tre occhi, dove l'occhio soprannumerario ha sul capo una posizione anomala. In questa rana la proiezione dalla retina soprannumeraria al tetto generalmente trasmette un'immagine dell'ambiente circostante che non concorda bene con quella proveniente dalla retina normale. Ciò significa che i neuroni vicini tra loro, ma in strie adiacenti in un lobo del tetto, ricevono informazioni su parti dello spazio visivo che non sono collegate tra loro. Si potrebbe simulare questa situazione adattando a uno dei nostri occhi un prisma che inclina la luce, per esempio di 90 gradi. Guardandosi attorno, il prisma trasmetterebbe in un occhio immagini dal cielo sopra di noi, mentre l'altro occhio vedrebbe il terreno di fronte a noi. Entrambi i tipi di immagini sarebbero segnalati simultaneamente alla stessa parte della corteccia visiva. Il mondo esterno rivestirebbe, quindi, scarso significato.

Una rana con tre occhi, di fronte alla quale viene fatto passare un oggetto che si muove irregolarmente e che imita la preda (ad esempio, un insetto che vola), rimane spesso immobile. Di tanto in tanto, cerca di colpire in modo aberrante lo stimolo. Se le viene concesso di vedere solo attraverso gli occhi normali i colpi che assesta sono precisi. Presumibilmente, le mappe nel tetto ottico che rappresentano gli occhi normali sono correttamente allineate con le vie motrici del sistema nervoso, che controllano il comportamento della rana. Se questa guarda solo con l'occhio soprannumerario, i colpi sono sempre diretti in modo sbagliato. Le vie motrici, che ricevono il comando dalla mappa visiva male allineata, fanno muovere il corpo dell'animale in direzioni che non corrispondono alla posizione della preda nello spazio.

Gli interrogativi sul significato funzionale o evolutivo delle strie nell'encefalo di rane con tre occhi sono chiaramente non pertinenti. Il terzo occhio è anomalo e, in mancanza di stimoli consistenti da ambedue gli occhi normali a un unico lobo del tetto ottico, la rana normale non trarrebbe beneficio da un meccanismo che si è evoluto in modo specifico per separare in strie gli stimoli che giungono al tetto. D'altra parte, la sopravvivenza delle rane libere e, in particolare, la loro capacità di catturare gli insetti di cui poi si nutrono dipendono in modo critico da un solido meccanismo, il quale assicura che in ogni lobo del tetto ottico si sviluppi una mappa esatta della superficie retinica controlaterale.

Pertanto, abbiamo cominciato a considerare la possibilità che anche le strie si possano formare grazie allo stesso meccanismo che genera le mappe. Un tale legame era stato suggerito per la prima volta fin dal 1975 da Simon LeVay, che lavorava in collaborazione con Hubel e Wiesel alla Harvard Medical School. LeVay ha avanzato la proposta che le strie funzionali nella corteccia visiva della scimmia rappresentino un compromesso tra due tendenze in conflitto: un processo di diffusione, in cui gli assoni che veicolano l'informazione da ciascuna delle due retine cercano di occupare l'intera corteccia visiva con una mappa, e un processo di raggruppamento in cui gli assoni che veicolano l'informazione da ciascuna delle retine cercano di rimanere uniti, come se fossero respinti dalle stimolazioni provenienti dall'altro occhio. Il risultato più probabile del conflitto tra queste due tendenze sarebbe la formazione di strie alternantisi, dato che una configurazione a strie ottimizzerebbe simultaneamente ambedue i processi.

Quali sono, allora, i meccanismi che danno origine alle mappe del sistema nervoso? Come potrebbero produrre strie quando due popolazioni di assoni formerebbero mappe di sé in un'unica zona bersaglio? Fortunatamente, le proiezioni dalla retina al tetto ottico nei vertebrati inferiori sono state studiate a lungo, indagando sulla formazione delle mappe del sistema nervoso e R. W. Sperry del California Institute of Technology è stato uno dei primi che se ne è occupato. Egli ha ruotato di 180 gradi, mediante intervento chirurgico, gli occhi dei tritoni. In alcuni soggetti ha lasciato intatti i nervi ottici; in altri ha sezionato tali nervi e quindi li ha fatti rigenerare. In entrambi i casi, i tritoni hanno compiuto degli errori di 180 gradi quando hanno cercato di afferrare la fonte dello stimolo e tali errori non si sono attenuati con il passare del tempo. Gli animali si comportavano come se non sapessero che le loro retine erano state ruotate. Evidentemente ogni parte della retina continuava a proiettare i propri assoni su una particolare porzione del tetto ottico, malgrado il fatto che Sperry fosse intervenuto per far sì che ogni settore della retina intercettasse settori anomali del mondo visibile. Numerosi studi successivi che hanno ampliato il lavoro di Sperry hanno dimostrato che la parte del tetto ottico che sarà innervata da una particolare porzione della retina è determinata nell'embrione già prima che gli assoni lascino la retina e si accrescano nell'encefalo in via di sviluppo.

Nel 1963, Sperry ha proposto una teoria per spiegare il coerente allineamento delle mappe visive nel cervello, suggerendo che le cellule della retina e le cellule del tetto ottico si sviluppino in funzione della loro posizione lungo ciascuno dei due assi, rispettivamente nella retina e nel tetto, cosicché ogni cellula viene ad avere sulla sua superficie una serie unica di molecole di marcatore. Gli assoni delle cellule retiniche possono, dunque, formare sinapsi solo con le cellule del tetto che portano i marcatori complementari. In breve, Sperry ha potuto visualizzare una corrispondenza piuttosto rigida tra retina e tetto, basata sull'affinità chimica. Questa corrispondenza non è tuttavia assoluta. Esperimenti compiuti in parecchi laboratori su pesci e anfibi hanno dimostrato che, in certe condizioni, gli assoni della retina formano sinapsi con assoni del tetto che non sono i loro normali bersagli. Una retina ridotta a metà da un intervento chirurgico può inviare i propri assoni, in modo che formino, attraverso un intero lobo del tetto ottico, una proiezione dilatata. Per contro, gli assoni che provengono da una retina intera possono comprimere la loro mappa, così che possa stare in mezzo tetto ottico, ridotto per intervento chirurgico.

È chiaro che un meccanismo basato su una corrispondenza rigida di marcatori fissi sulle cellule della retina e su quelle del tetto ottico non è in grado di spiegare la plasticità che viene messa in evidenza quando si alterano chirurgicamente le dimensioni delle popolazioni di cellule retiniche e di cellule del tetto. Invece, le

posizioni degli assoni retinici nel tetto devono essere controllate da qualche meccanismo che può adeguarsi ai cambiamenti dei numeri relativi delle cellule retiniche e del tetto.

In che modo si realizza tale adeguamento? Sono state formulate tre possibilità. In primo luogo, i marcatori con chemoaffinità rigida tra loro, proposti da Sperry, potrebbero essere capaci di una «rispecificazione», cosicché la perturbazione provocata dall'intervento chirurgico potrebbe far modificare il marcatore nella retina o nel tetto. In secondo luogo, l'identificazione delle cellule nella retina o nel tetto, invece di dipendere da molti marcatori differenti, potrebbe dipendere da gradienti, uno lungo ognuno dei due assi, delle due sostanze marcatrici.

La terza possibilità è che la retina e il tetto non abbiano affatto marcatori. Gli assoni che si proiettano dalla retina al tetto potrebbero mantenere il loro ordinamento relativo, per mezzo di una coesione che conservano reciprocamente man mano che crescono verso il tetto. Si deve allora spiegare in che modo la mappa come insieme venga sempre ad avere, nel tetto, lo stesso orientamento. In particolare, le mappe della retina in tutti i vertebrati non mammiferi rappresentano la parte centrale del mondo visibile dell'animale nella parte anteriore del lobo del tetto ottico, e le parti più laterali nelle parti posteriori dello stesso lobo.

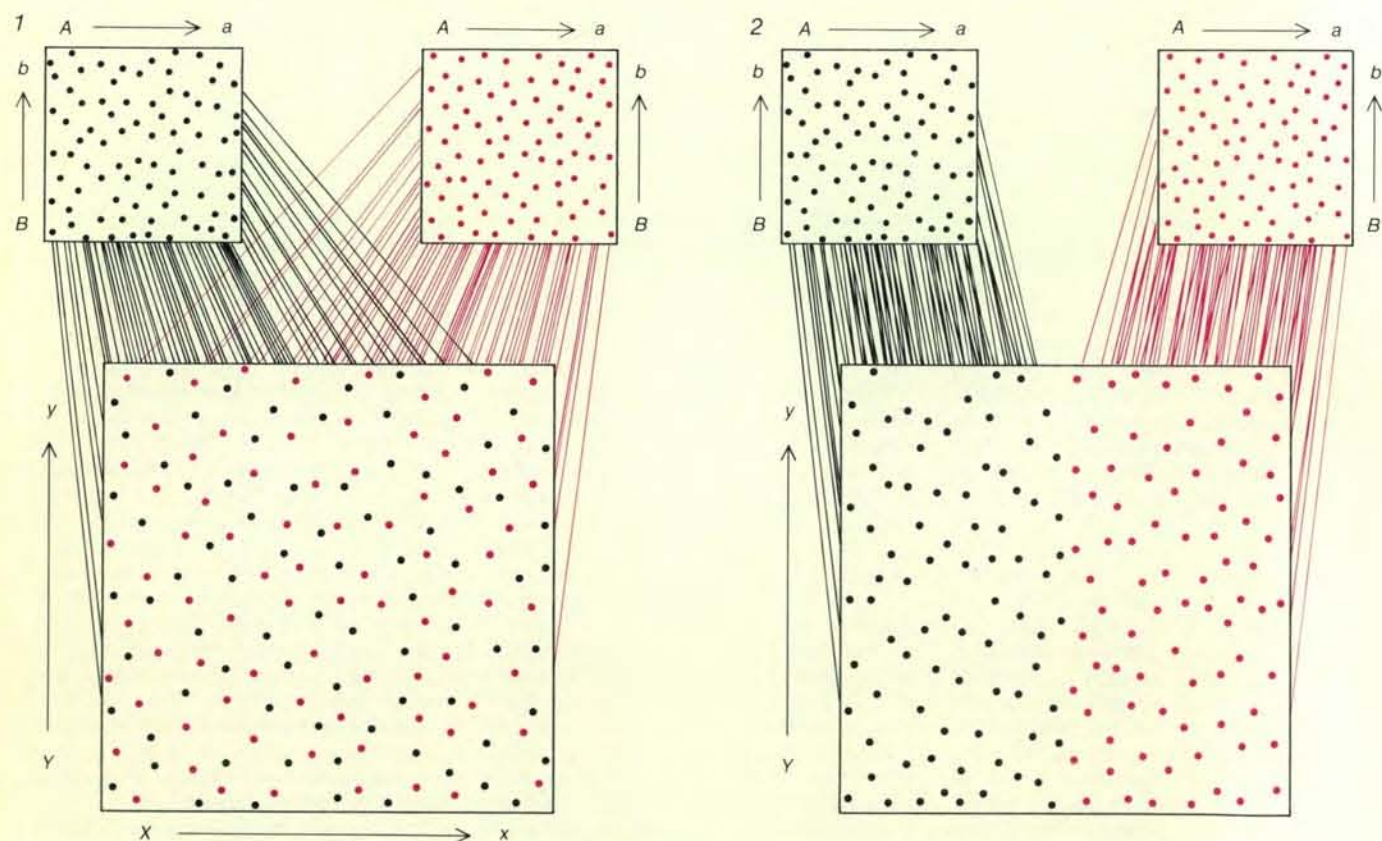
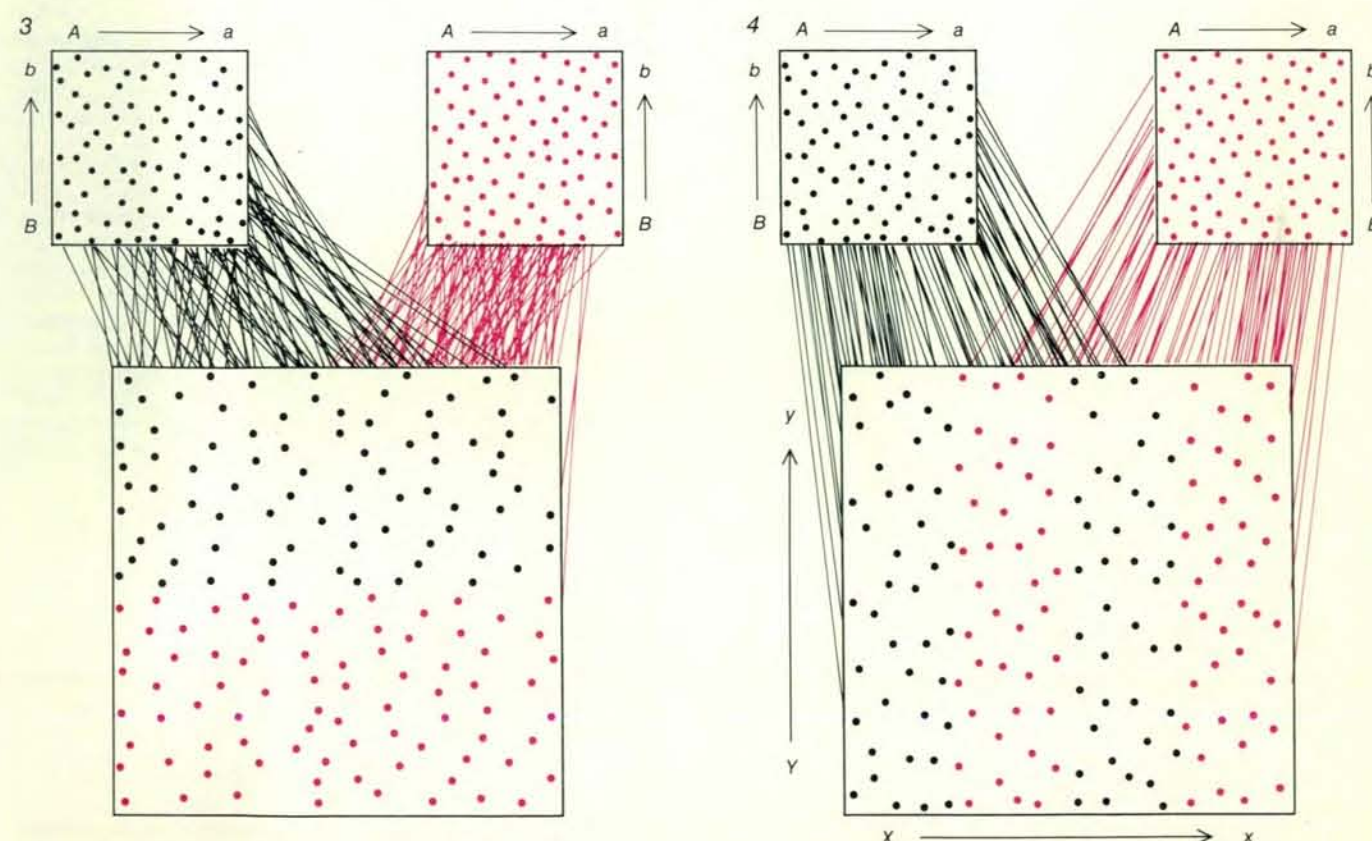
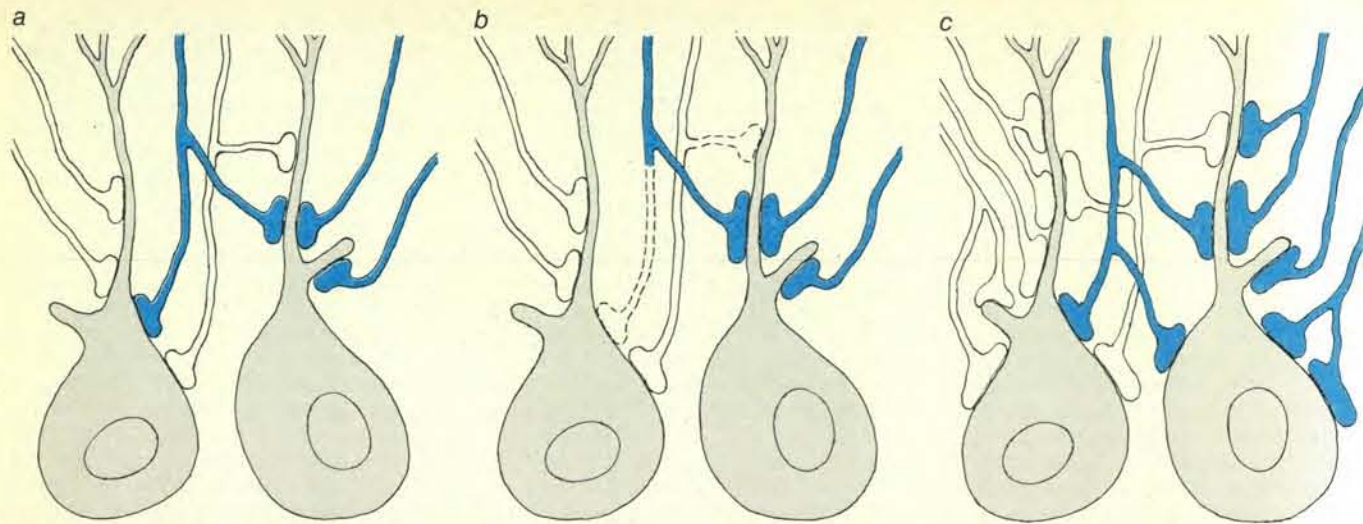
Parecchi indizi sono disponibili oggi per meglio valutare queste varie possibilità. Di fatto, la prima possibilità, cioè l'idea di una modificazione o rispecificazione dei marcatori rigidi della retina o del tetto, può non essere valida. Da una parte, una serie di manipolazioni chirurgiche, realizzate sul carassio, ha dimostrato che un tetto ottico può ricevere in sequenza stimoli da una retina normale, quindi stimoli da una mezza retina dilatata, quindi di nuovo stimoli da una retina normale. Si hanno inoltre alcune descrizioni di esperimenti su anuri del genere *Xenopus*, in cui una regione del tetto (anche se, probabilmente, non le stesse cellule del tetto) riceve simultaneamente stimoli dalla metà di una retina, anteriore dal punto di vista embrionale, e dalla metà dell'altra retina, posteriore dal punto di vista embrionale. Così, se i marcatori del tetto rispecificano, sono in grado di farlo frequentemente. Inoltre, le cellule all'interno di una piccola regione del tetto possono modificare i loro marcatori indipendentemente dalle loro vicine. I marcatori del tetto devono essere dotati di una tale plasticità che non è possibile identificare una cellula in base alla sua posizione nel tetto.

Da parte loro, i marcatori nella retina, se esistono, non sembra che si modifichino. Scott E. Fraser, che lavora alla Johns Hopkins University, ha asportato un occhio da girini dell'anuro *Xenopus*. L'occhio rimasto intatto ha quindi proiettato degli assoni sul lobo controlaterale del tetto ottico, proprio come avrebbe fatto in condizioni normali. Inoltre, la parte ventrale (inferiore) dell'occhio intatto, il cui apporto al nervo ottico si stava ancora sviluppando nel momento dell'intervento chirurgico, ha inviato assoni a tutto il lobo ottico dello stesso lato del tetto ottico, cioè a quel lobo che sarebbe stato innervato dall'occhio asportato. Se una regione ventrale di una retina potesse proiettare una mappa dilatata a un lobo del tetto e una mappa normale all'altro lobo, essa dovrebbe connettersi con cellule in differenti posizioni all'interno del tetto, il che rende improbabile che la dilatazione di una mappa comporti la rispecificazione dei marcatori della retina.

La seconda possibilità, quella dei marcatori graduati nella retina e nel tetto, spiega le dilatazioni o le compressioni di una mappa. Essa spiega la capacità delle cellule di una determinata regione del tetto di ricevere gli assoni da differenti parti di una retina; spiega anche la facilità con cui una parte di una retina può inviare assoni a differenti parti di due lobi del tetto. Un gradiente separato di una molecola di marcatore lungo ciascuno dei due assi è sufficiente a fornire a ogni posizione nella retina e nel tetto una combinazione unica di marcatori e la corrispondenza tra le posizioni nella retina e le posizioni nel tetto può adattarsi alla gamma dei marcatori presenti nella retina o nel tetto.

L'ipotesi dei marcatori graduati prevede, però, che l'orientamento di una proiezione della retina sul tetto si conservi anche dopo una qualsiasi perturbazione. Alcuni esperimenti dimostrano il contrario. Ronald L. Meyer, che lavora al California Institute of Technology, ha eliminato in un carassio la metà di una retina, operazione che ha ridotto a metà gli stimoli a uno dei due lobi del tetto ottico. Nel contempo, egli ha costretto la metà degli assoni provenienti dall'altro occhio a portarsi nella metà del lobo del tetto ottico rimasta vacante. Si poteva prevedere che gli assoni deviati sarebbero terminati nel lobo del tetto ottico rimasto a metà libero praticamente allo stesso modo in cui sarebbero terminati nel lobo sul quale avrebbero dovuto proiettarsi in condizioni normali. In questo esperimento, gli assoni che Meyer aveva deviato avrebbero dovuto terminare nella parte del lobo del tetto ottico rimasto libero a metà e che aveva conservato la stimolazione proveniente dalla retina. Al contrario, questi assoni deviati hanno formato una proiezione erroneamente orientata (di fatto invertita) nella metà vacante del lobo del tetto ottico. È evidente che essi, dal centro dell'occhio intatto, hanno raggiunto, nel tetto ottico, una posizione quanto più possibile vicina a quella giusta. Gli altri assoni deviati, invece, non hanno potuto conservare né la continuità né l'orientamento della loro mappa retinica, in quanto sarebbero stati costretti a terminare nella parte del lobo del tetto ottico già occupata. Lo studio di Meyer indica così che la conservazione dei rapporti di vicinato è una tendenza importante, che può operare indipendentemente nella costituzione di una mappa. Nel suo esperimento, in fondo, i rapporti di vicinato erano stati mantenuti in una regione del tetto ottico non giusta e a spese del normale orientamento della mappa.

Risultati di questo genere sembrano sostenere la terza possibilità, che favorisce una coesione tra gli assoni che si protendono verso il tetto ottico e propone che non vi sia tra le cellule del tetto e le cellule della retina una corrispondenza per affinità chimica. Chi ha proposto quest'idea cita studi che indicano come nei pesci, negli anuri e nei polli gli assoni provenienti da molte parti (non da tutte) della retina si protendono verso il tetto assieme ad assoni provenienti da cellule che sono loro vicine nella retina. Come abbiamo già notato, però, l'idea non spiega perché le mappe normali sono orientate in modo coerente. La mancanza di marcatori della retina e del tetto è anch'essa difficile da conciliare con un vasto numero di esperimenti, in cui gli assoni retinici interrompono la continuità della loro mappa per andare a finire opportunamente in un pezzo del tessuto che costituisce il tetto e che è stato ruotato o trapiantato in una posizione anomala del tetto stesso.

Ciò che più di tutto ci ha convinto dell'esistenza dei marcatori della retina e del tetto ottico è stato un risultato che abbiamo ottenuto con rane a tre occhi. In circa un quarto di queste, la retina soprannumeraria ha inviato assoni ad ambedue i lati dell'encefalo; così né l'uno n'é l'altro lobo del tetto ottico si presentava del tutto striato. In queste rane, un buco nella distribuzione delle strie in uno dei due lobi è risultato corrispondente a un insieme di strie, localizzate nel punto specularmente simmetrico dell'altro lobo. Se gli assoni del terzo occhio, che si prolungano verso l'interno, semplicemente conservano la topologia della retina, le loro

Vengono qui messi a confronto alcuni possibili meccanismi mediante i quali si formano le mappe cerebrali. In alto, sopra ogni parte dell'illustrazione, sono state rappresentate le due retine. Le loro cellule *(punti)* sarebbero marcate da gradienti nella concentrazione di due sostanze *(A e B)* alla superficie. In basso, in ogni parte dell'illustrazione, si trova un lobo del tetto ottico, a cui le rispettive retine inviano i loro assoni. Nella forma più semplice del meccanismo chiamato «corrispondenza per affinità chimica» *(1)* si ritiene che le cellule presenti nel lobo siano marcate da gradienti *(X, Y)*, la cui complementarietà nei riguardi dei marcatori della retina guida gli assoni che si stanno sviluppando. Gli assoni di ambedue gli occhi si mescolano non appena le loro terminazioni raggiungono il tetto ottico, un fatto che non è mai stato rilevato nelle rane con tre occhi. Secondo un altro meccanismo ipotizzabile, leggermente diverso *(2)*, gli assoni di ogni retina mantengono il loro ordine spaziale sviluppandosi verso il tetto. Per orientare la mappa, il lobo fornisce solo un'informazione sufficiente (in questo caso un solo gradiente). Ogni retina innerva una regione separata del lobo, una situazione che in realtà non si riscontra mai. In base a un altro possibile meccanismo *(3)*, gli assoni mantengono il loro ordine, ma non ricevono alcuna informazione dal tetto. In questo caso producono mappe ruotate. Anche questa situazione non la si riscontra mai nella realtà. In un quarto meccanismo *(4)* operano due processi. In primo luogo un accostamento per una imprecisata chemoaffinità disperde le terminazioni assoniche sul lobo del tetto ottico secondo un corretto orientamento. Quindi una serie di interazioni locali mantiene vicine nel lobo solo quelle terminazioni assoniche che provengono da cellule che sono vicine in una delle due retine. Ne risultano delle strie, perché solo esse rendono ottimale ciascuno dei due processi, simultaneamente l'uno all'altro.

Gli eventi cellulari che si pensa siano alla base dello sviluppo di una mappa precisa in un lobo del tetto ottico sono schematizzati qui per due cellule di un lobo innervato da ambedue gli occhi. Un episodio iniziale di corrispondenza per affinità chimica produce delle sovrapposizioni locali in cui assoni provenienti dai due occhi convergono sulle cellule dello stesso lobo *(a)*. Quindi le terminazioni che rappresentano un occhio sono rafforzate (nel disegno si suppone che diventino più grandi) a spese delle terminazioni che rappresentano l'altro occhio *(b)*. Inoltre, la densità delle terminazioni «corrette» potrebbe aumentare *(c)*. In ambedue i casi ogni retina finisce per dominare gruppi di cellule. Le terminazioni che rappresentano cellule vicine in una retina trasmettono con tutta probabilità segnali ben correlati. Questa correlazione potrebbe essere alla base del rafforzamento di connessioni di cui esistono mappe precise e, inoltre, potrebbe essere responsabile delle strie.

proiezioni verso i lobi del tetto ottico avrebbero dovuto espandersi e formare strie in ambedue i lobi. È chiaro, invece, che gli assoni soprannumerari possono competere per lo spazio del tetto ottico con gli assoni provenienti dalla retina normale solo in punti adatti per la parte di retina da cui gli assoni emergono. Sembra, dunque, che le cellule del tetto siano marcate e che gli assoni retinici siano in grado di discriminare tra i vari marcatori.

Recenti ricerche condotte nel nostro laboratorio mostrano, inoltre, che un lobo del tetto ottico mai innervato da una retina può comunque sviluppare una mappa. Abbiamo asportato da embrioni di rana ambedue i primordi degli occhi, molto prima che essi cominciassero a inviare assoni all'encefalo. In seguito, abbiamo cercato di individuare le proiezioni assoniche con cui il tetto ottico aveva potuto stabilire mappe in altre parti dell'encefalo. Queste erano comunque identiche nelle rane normali e in quelle senza occhi. Pertanto, le cellule del tetto ottico sono in grado di esprimere le loro identità di posizione indipendentemente dalle loro connessioni con la retina. È chiaro che nei lobi del tetto ottico deve essere disponibile una qualche forma di informazione, che possa assicurare una buona messa a registro della mappa retinica con altre mappe visive nell'encefalo.

Sembra, così, ovvio che nessun meccanismo semplice di formazione delle mappe riuscirà a risolvere le controversie che emergono da molte osservazioni sperimentali. Da una parte, sembra che gli assoni di una certa porzione dell'occhio siano in grado di trovare una certa parte di un lobo del tetto ottico. Dall'altro, numerosi studi (e le strie nel lobo del tetto ottico delle rane a tre occhi, che presenta una duplice innervazione) riflettono una capacità di coesione tra le terminazioni sinaptiche, che provengono da una retina, capacità che non può essere spiegata da una corrispondenza tra retina e tetto ottico, dovuta ad affinità chimica.

Se si ammette, tuttavia, che due meccanismi indipendenti operino nella definizione delle mappe nervose, molte controversie scompaiono. Inoltre, la formazione di strie diventa una logica estensione della formazione di mappe. Si supponga che, in una fase precoce dello sviluppo di una mappa, delle chemoaffinità si siano distribuite con una certa gradazione lungo almeno due assi della retina e che il tetto ottico abbia guidato lo sviluppo degli assoni. Non è necessario che questa guida sia precisa: i gradienti possono anche essere superficiali e le affinità deboli. Nel sistema visivo della rana leopardo è necessario solo ammettere che ogni assone arrivi nel giusto quadrante del lobo del tetto ottico. La precisione della mappa risulterebbe da un secondo stadio di sviluppo, in cui interazioni nell'ambito del lobo del tetto manterrebbero come vicine solo quelle terminazioni assoniche che hanno origine da cellule che sono vicine nella retina. Il risultato di una simile sequenza sarà il compromesso riconosciuto da Le Vay, Hubel e Wiesel, in cui la zona bersaglio delle due proiezioni è suddivisa in bande terminali allungate.

Il fascino dell'ipotesi della chemoaffinità ha ispirato i ricercatori a indagare sulle molecole dei marcatori presenti sulla superficie delle cellule della retina e del tetto ottico. Tali molecole devono essere distribuite nella retina o nel tetto ottico con un gradiente e con una capacità di legare altre sostanze, che potrebbero dare origine al ben noto allineamento della mappa nel tetto ottico. Parecchi recenti progressi fanno sperare in un successo. Per esempio, alcuni ricercatori nel laboratorio di Marshall W. Nirenberg, presso il National Heart, Lung and Blood Institute, espongono cellule del sistema immunitario del topo a estratti di retina di pollo. Quindi isolano e clonano le cellule della milza del topo, che producono anticorpi. Ogni coltura cellulare risultante sintetizza un anticorpo estremamente specifico, e uno degli anticorpi ottenuti in questo modo appare in grado di legarsi in maniera graduata con cellule disposte lungo un asse della retina. Ne segue che la molecola ignota a cui l'anticorpo si lega ha una analoga distribuzione graduata. Con una diversa impostazione sperimentale, Willi Halfter, Michael Claviez e Uli Schwartz del Max Planck Institut für Virusforschung di Tübingen hanno posto la questione dell'adesione tra retina e tetto. Essi trovano che gli assoni di differenti parti della retina di pollo mostrano differenze consistenti nella capacità di legare tra loro membrane isolate da cellule di tetto ottico di pollo.

Questioni di fondo continuano, comunque, ad assediare il secondo stadio di formazione delle mappe: si tratta dei problemi relativi alle interazioni che tengono unite le terminazioni provenienti da cellule vicine e presumibilmente danno origine a strie. Michael P. Stryker della School of Medicine dell'Università della California di San Francisco ha mostrato che la tetrodotossina, una sostanza chimica che blocca la capacità dei neuroni di trasmettersi reciprocamente segnalazioni mediante picchi di tensione, chiamati potenziali d'azione, impedisce o ritarda lo sviluppo delle strie nella corteccia visiva, se viene iniettata negli occhi di un gattino. Le proiezioni degli occhi rimangono mescolate nella zona corticale che di tali proiezioni costituisce il bersaglio. N. V. Swindale dell'Università di Cambridge ha riportato analoghi risultati con gattini allevati al buio.

È evidente che l'attività nervosa è essenziale per la coesione tra le terminazioni sinaptiche che rappresentano l'uno o l'altro occhio. Come interverrebbe? All'interno di una data retina, cellule vicine che proiettano i loro assoni in direzione del tetto ottico (o verso la corteccia visiva) tendono a generare analoghe sequenze di potenziali d'azione in quanto sono connesse (mediante neuroni retinici intermedi) con molte delle stesse cellule fotorecettrici. Inoltre, i potenziali d'azione correlati dei neuroni che sono vicini nella retina inducono, con maggiore probabilità dei segnali non correlati, un'attività elettrica in una determinata cellula del tetto ottico. Pertanto, l'attività ben correlata in corrispondenza di coppie di sinapsi potrebbe presumibilmente servire, nel tetto (o nella corteccia visiva), a marcare le sinapsi di cellule che sono vicine nella retina. Se i neuroni del tetto ottico avessero il compito di rinforzare le sinapsi provenienti da parecchi neuroni ben correlati, a spese di sinapsi la cui emissione di segnali fosse relativamente poco efficace, una mappa topografica grossolana diventerebbe anche una mappa precisa.

In breve, un modello del modo in cui si sviluppa la mappa del tetto ottico, modello che si basa su due meccanismi, propone l'esistenza di deboli affinità graduate che conducono gli assoni retinici nel tetto ottico a un grossolano allineamento. Tale mappa viene quindi ordinata con precisione dal rafforzamento delle sinapsi che provengono da cellule retiniche vicine, le quali tendono a essere simultaneamente attive. Cristoph von der Malsberg e David Willshaw del Max Planck Institut für biophysikalische Chemie di Göttingen hanno ideato delle simulazioni al calcolatore in cui il rafforzamento selettivo delle sinapsi opera su due proiezioni topografiche grossolane in un'unica zona bersaglio. Essi hanno trovato che le simulazioni danno origine a mappe con strie.

L'idea che l'efficienza delle terminazioni sinaptiche possa determinarne la stabilità e la posizione in seno all'encefalo non è né nuova né limitata alle mappe. Negli anni quaranta, D. O. Hebb della McGill University ha suggerito che il rafforzamento selettivo delle sinapsi potesse essere alla base di certi aspetti dell'apprendimento. Varianti dell'idea di Hebb sono state espresse da allora per spiegare lo sviluppo dei collegamenti nervosi nel cervelletto, la sensibilità dei neuroni sensoriali a particolari stimoli, e la maturazione delle connessioni motorie tra il sistema nervoso e i muscoli. Benché i neurologi siano oggi ancora molto lontani dalla spiegazione dei meccanismi molecolari che sarebbero alla base del rafforzamento selettivo o della stabilizzazione delle sinapsi, il concetto in sé è utile per tentare di capire come l'attività nervosa può influenzare la struttura nervosa. I potenziali d'azione e la relativa efficacia dei segnali sinaptici sono probabilmente il legame che unisce le mappe e le strie nell'encefalo. Potrebbero, in effetti, armonizzare bene il sistema nervoso che si sta sviluppando.

Campioni di Via Lattea

L'osservazione da satellite dei raggi cosmici, ossia nuclei provenienti da altre regioni della nostra galassia, indica che la composizione isotopica di quelle regioni è diversa da quella del sistema solare

di Richard A. Mewaldt, Edward C. Stone e Mark E. Wiedenbeck

I progressi compiuti nella tecnologia spaziale e nella relativa strumentazione rendono ora possibile determinare la composizione di campioni di materia provenienti da qualunque regione della nostra galassia, la Via Lattea. Sebbene il campionamento diretto di materia proveniente da regioni situate al di fuori del sistema solare sia fuori questione, si può compiere la stessa operazione intercettando autentici «messaggeri stellari», ossia i raggi cosmici che sono frammenti di materia a elevata energia provenienti da altre stelle. Analisi di raggi cosmici compiute negli ultimi quattro anni mediante gli strumenti degli osservatori specializzati a bordo di satelliti rivelano che la composizione dei raggi cosmici provenienti dalla nostra galassia è nettamente diversa dalla composizione del Sole e di altri corpi del sistema solare. È probabile che tali differenze di composizione derivino dalle diverse condizioni nelle quali gli elementi sono sintetizzati all'interno delle stelle.

La nostra galassia è costituita per il 90 per cento circa da idrogeno, per l'1 per cento da elementi più pesanti e per il rimanente 9 per cento da elio. Le stelle brillano soprattutto per la conversione termonucleare dell'idrogeno in elio e dell'elio in elementi più pesanti, con una conseguente liberazione di energia. La composizione di una galassia muta col tempo, man mano che alcuni nuclei atomici sintetizzati all'interno di stelle massive vengono espulsi da esplosioni di supernove nel gas interstellare, il quale pertanto accumula gradatamente atomi provenienti da molte stelle. Il sistema solare rappresenta un campione della miscela di atomi che condensarono circa 5 miliardi di anni fa dal gas interstellare in una minuscola regione della Galassia. Può darsi che la miscela non sia tipica della Galassia nel suo insieme.

Alcuni nuclei espulsi da supernove possono essere accelerati durante l'esplosione fino a raggiungere quasi la velocità della luce e a diventare raggi cosmici. Altri nuclei ancora del gas interstellare possono essere accelerati a elevate velocità dalle onde d'urto causate da supernove più recenti. Poiché i nuclei atomici sono elettricamente carichi, interagiscono con il campo magnetico della galassia; essi vengono imprigionati dal campo magnetico per periodi di tempo dell'ordine di 10 milioni di anni prima che possano alla fine sfuggire nello spazio intergalattico. Prima che i raggi cosmici lascino la Galassia, alcuni di essi passano vicino alla Terra, dove possono essere catturati e analizzati. Tuttavia la cattura e l'analisi devono essere compiute nello spazio, poiché i raggi cosmici che entrano nell'atmosfera terrestre vengono disintegrati per collisione con i nuclei degli atomi presenti nell'atmosfera stessa, prima che possano raggiungere la superficie terrestre.

Le prime ricerche compiute con strumenti trasportati da palloni d'alta quota o installati su navi spaziali hanno stabilito che l'1 per cento circa dei raggi cosmici è costituito da nuclei di elementi più pesanti dell'idrogeno e dell'elio. Queste ricerche hanno dimostrato anche che nei raggi cosmici l'abbondanza relativa dei comuni elementi più pesanti (come carbonio, ossigeno, magnesio, silicio e ferro) è circa uguale a quella presente nel sistema solare. Tuttavia migliorando la precisione delle misurazioni dei raggi cosmici si può indagare non solo sulla abbondanza dei nuclei atomici, ma anche su quella di isotopi; ossia delle specie di nuclei atomici.

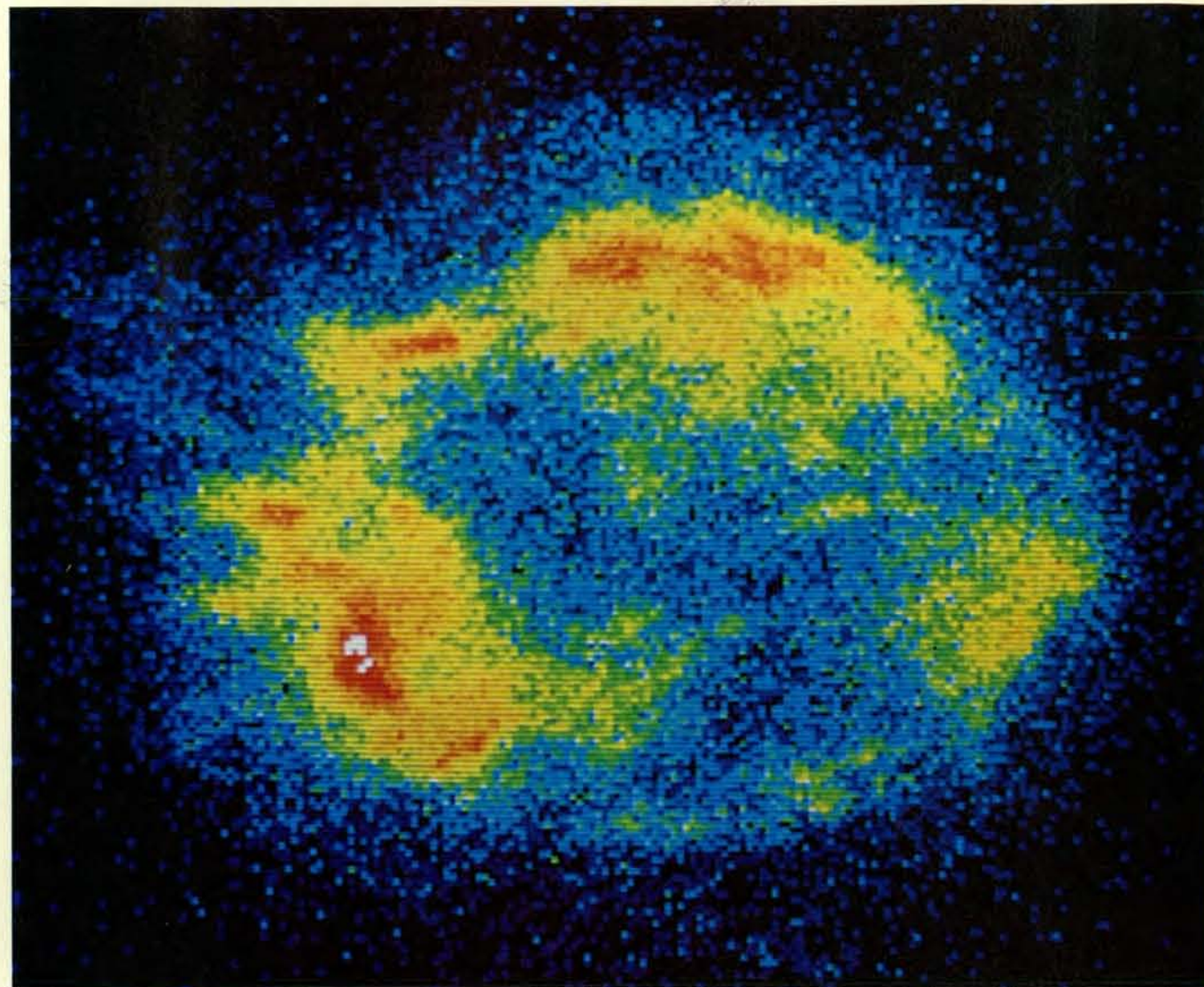
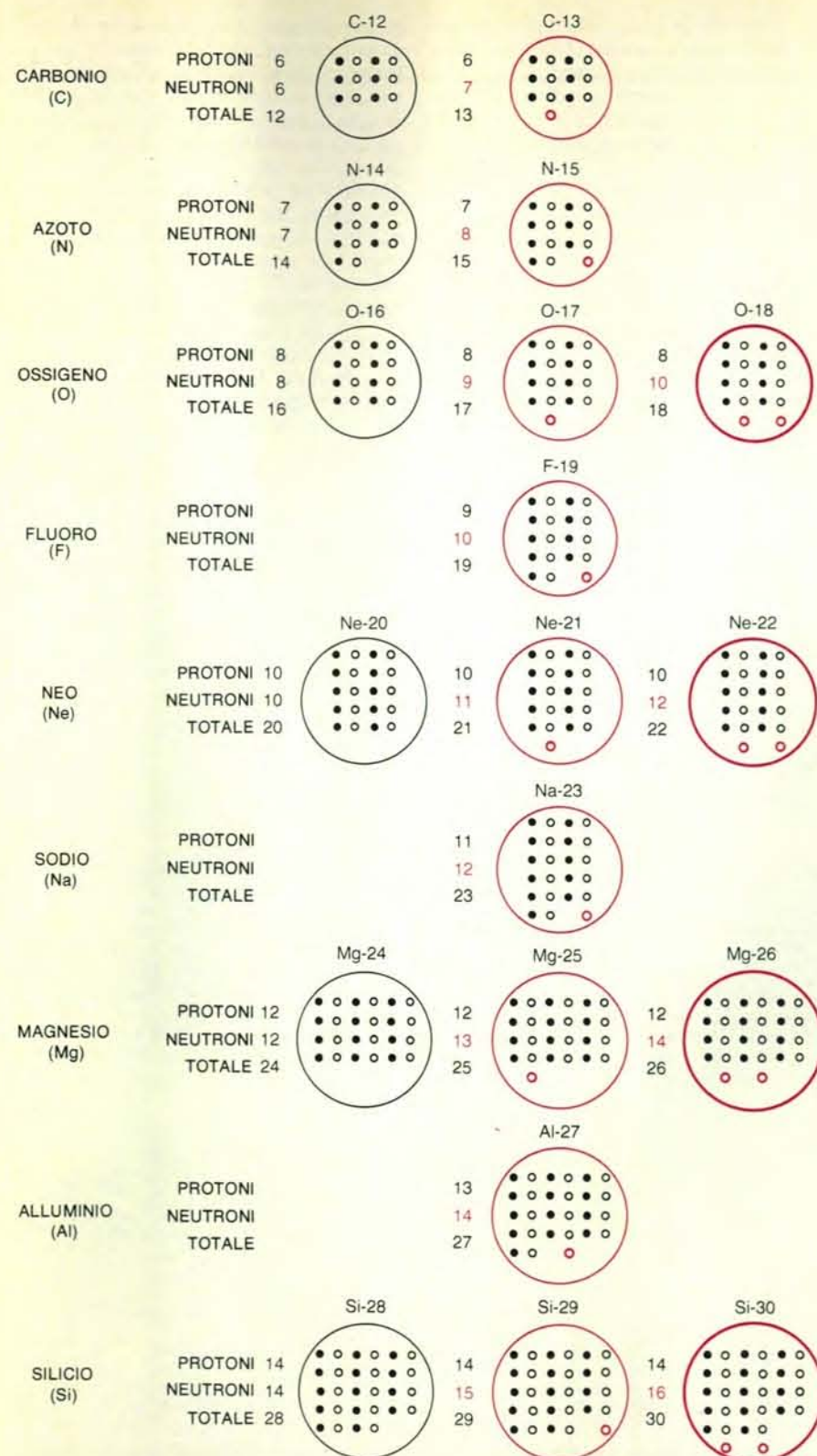
Gli isotopi degli elementi si differenziano per la loro massa. Gli elementi differiscono fra di loro per il numero di protoni contenuti nel nucleo (tale numero viene indicato con Z): dall'idrogeno che contiene un protone singolo si arriva via via fino all'uranio che ne contiene 92. Gli isotopi di un elemento differiscono per il numero di neutroni contenuti nel nucleo: nessun neutrone nell'idrogeno comune (H-1), uno nell'idrogeno pesante, o deuterio (H-2), fino a 143 e 146 neutroni nei due isotopi più abbondanti dell'uranio, U-235 e U-238. Poiché i protoni e i neutroni hanno circa la stessa massa, gli isotopi di un elemento differiscono fra di loro per numeri interi di unità di massa. Misurando le abbondanze isotopiche nei nuclei di raggi cosmici si può sperare non solo di scoprire una materia che può avere una composizione diversa da quella del sistema solare, e pertanto una storia diversa, ma anche di saperne di più sui differenti processi di costituzione degli elementi, processi che hanno luogo all'interno di stelle di tipo diverso.

In laboratorio si può misurare facilmente la massa degli isotopi mediante lo spettrometro di massa, uno strumento di dimensioni e peso non proprio trascurabili. È più difficile effettuare la stessa operazione nello spazio, con uno strumento che pesi meno di 10 chilogrammi e il cui consumo di energia sia inferiore a 10 watt. Inoltre, lo strumento impiegato nello spazio deve lavorare con nuclei singoli a elevata energia e non con un campione di una certa quantità di materia. Oltre a ciò deve essere in grado di rimandare misure di elevata precisione mentre opera autonomamente per un periodo di anni. Due strumenti idonei furono costruiti, uno all'Università della California e l'altro al California Institute of Technology. Essi furono portati nello spazio nell'agosto 1978 dal Third International Sun-Earth Explorer (ISEE-3) della NASA, il primo satellite che trasportava strumenti progettati appositamente per misurare la composizione isotopica di nuclei pesanti di raggi cosmici. Il veicolo spaziale esegue altri dieci esperimenti per misurare molti altri fenomeni interplanetari e astrofisici.

L'ISEE-3 fu posto in un'insolita orbita distante circa 1,5 milioni di chilometri dalla Terra (un centesimo della distanza dal Sole) vicino al «punto di Lagrange» dove è soggetto ad attrazione gravitazionale sia da parte della Terra, sia da parte del Sole. Con l'aiuto di piccole spinte di gas l'ISEE-3 segue un percorso leggermente sinusoidale che lo porta lievemente al di sopra e poi al di sotto del piano dell'orbita terrestre. Quando attraversa il piano orbitale, si trova alternativamente un po' davanti o un po' dietro a una linea congiungente la Terra e il Sole. Pertanto, visto dalla Terra, l'ISEE-3 sembra eseguire un alone attorno al Sole, spostato da questo di un angolo di circa 15 gradi.

Lo scopo dei due spettrometri posti sul veicolo spaziale è quello di determinare la distribuzione della massa dei nuclei di raggi cosmici «pesandoli» uno per volta con una precisione superiore all'1 per cento. Il principio di funzionamento di entrambi gli strumenti comporta la misurazione dell'energia cinetica, E, e della velocità, v, di un nucleo in arrivo, per poter ricavare la massa, m, secondo l'equazione dell'energia cinetica: $E = 1/2\,mv^2$. Pertanto m è uguale a $2E/v^2$, dove la massa è espressa in unità di massa nucleonica, ossia la massa di un protone o di un neutrone. Vengono controllati nuclei che viaggiano a velocità comprese tra il 25 e il 75 per cento della velocità della luce. Essendo richiesta una precisione dell'1 per cento, è necessario misurare la velocità dei nuclei con una precisione di pochi decimi di per cento.

Gli elementi sensibili di entrambi gli strumenti sono rilevatori allo stato solido: sottili wafer di silicio, la cui superficie è di circa 10 centimetri quadrati. Quando un nucleo carico di energia passa attraverso questo wafer la carica elettrica positiva del nucleo libera alcuni elettroni dal reticolo cristallino del silicio. Il nucleo cede energia cinetica agli elettroni e così diminuisce la sua velocità. Applicando un potenziale tra gli elettrodi posti sulle superfici del rilevatore, è possibile raccogliere gli elettroni liberati, generando così un piccolo segnale elettrico proporzionale alla energia cinetica persa dal nucleo. Maggiore è il numero di protoni nel nucleo, più grande è la carica positiva e maggiore la quantità di energia che sarà ceduta agli elettroni. D'altra parte, più alta è la velocità del nucleo, minore è l'energia che perderà poiché il nucleo impiega meno tempo a passare attraverso il wafer di silicio. La perdita complessiva di energia da parte del nucleo nel passare attraverso un rilevatore è grosso modo proporzionale a Z^2/v^2, ossia al rapporto fra il quadrato della carica e il quadrato della velocità. Man mano che il nucleo passa attraverso rilevatori successivi, continua a rallentare e a perdere energia finché si ferma. La somma dei segnali elettrici provenienti dalla pila di rilevatori è una misura dell'energia cinetica posseduta in origine dal nucleo.

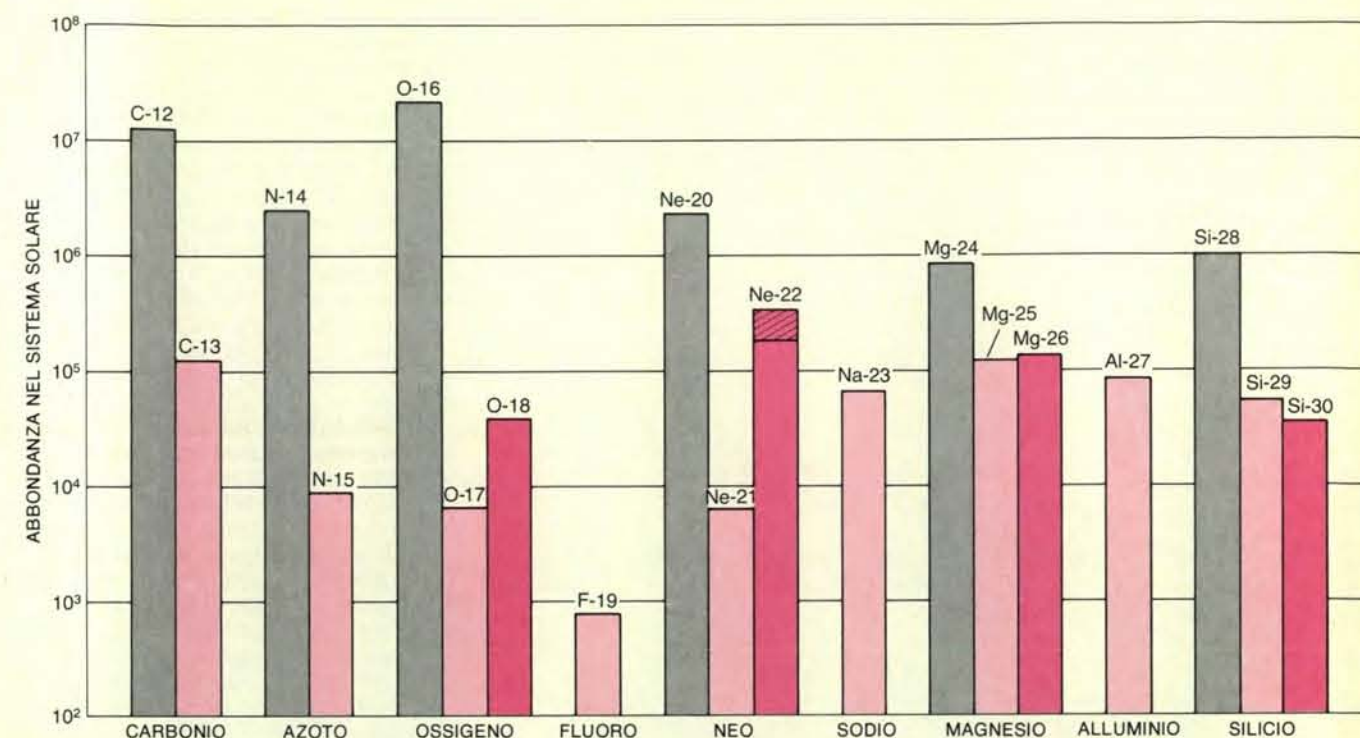
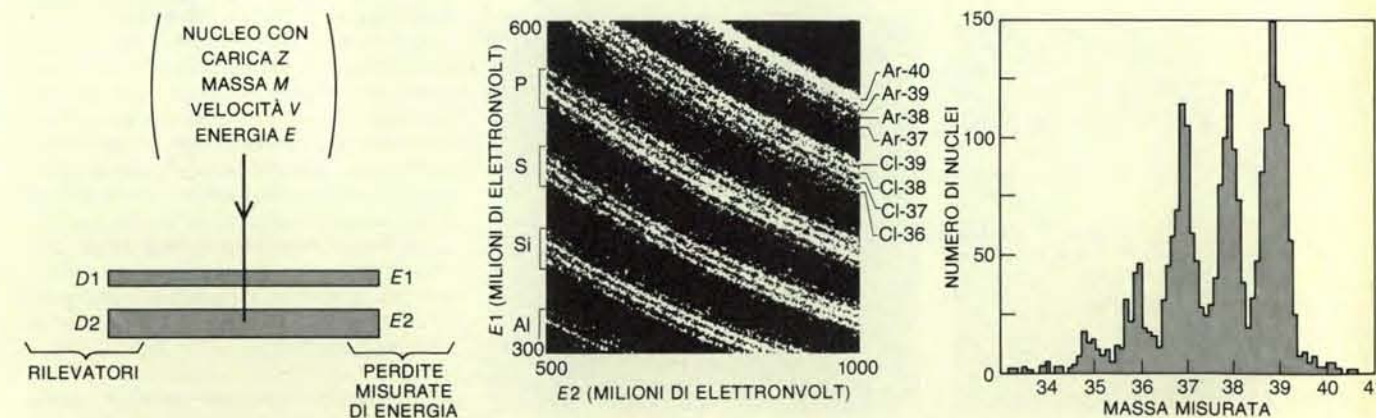
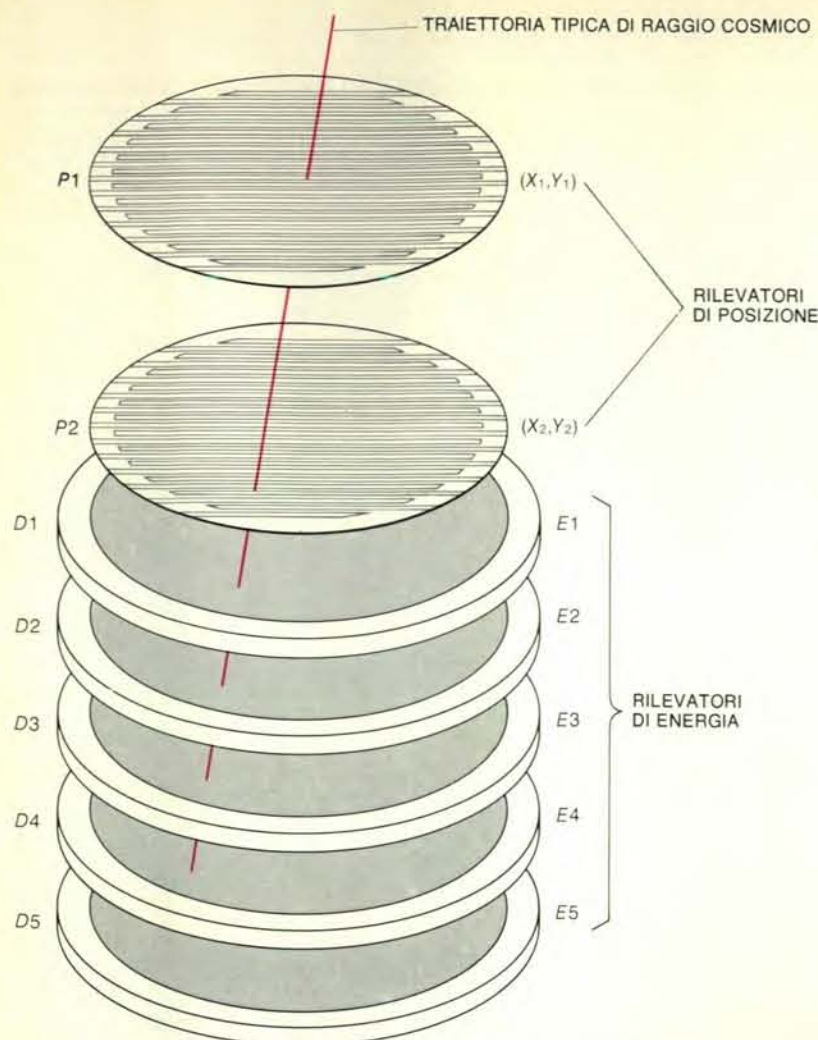
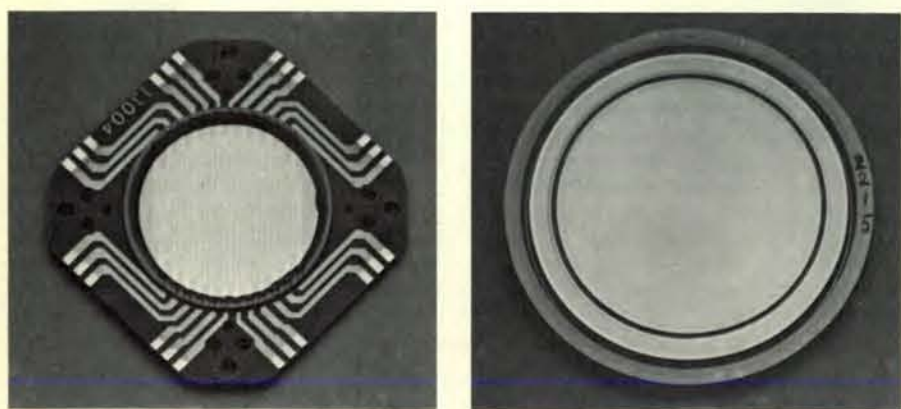
In teoria Z, m ed E possono essere determinati unicamente per singoli nuclei di raggi cosmici che vengono a fermarsi nello spettrometro. Tuttavia, l'energia persa nello spettrometro dipende anche dalla lunghezza del percorso del nucleo nello strumento, che varia naturalmente

Resti di supernova nella costellazione Cassiopeia, Cas A, sono un esempio di regione della Galassia in cui particelle nucleari vengono accelerate e diventano raggi cosmici. In questa immagine a raggi X a colori codificati (ottenuta con un rilevatore installato sull'Osservatorio Einstein) l'anello luminoso a intensa emissione di raggi X *(rosso e giallo)* è formato da un involucro in espansione di materiale stellare emesso da un'esplosione di supernova mista a gas interstellare. Un involucro di materia più grande *(blu)* è forse gas interstellare riscaldato dal fronte d'urto, diretto verso l'esterno, della esplosione. Si pensa che tali onde d'urto accelerino i raggi cosmici. Sulla base della distanza stimata di Cas A, di 9000 anni luce, l'anello luminoso ha un diametro di circa 9 anni luce. Dalle dimensioni e dalla velocità di espansione dell'anello, si calcola che l'esplosione sia avvenuta circa 325 anni fa, ma non c'è prova che sia stata osservata sulla Terra. Questi sono i più giovani resti di supernova conosciuti nella Galassia. Nelle esplosioni di supernova, gli elementi pesanti formatisi in stelle massive sono dispersi nel gas interstellare e fanno variare la composizione della Galassia nell'arco di un periodo di tempo. L'immagine a raggi X è stata ottenuta da S. S. Murray e collaboratori al Center for Astrophysics dello Harvard College Observatory e allo Smithsonian Astrophysical Observatory.

Sono qui illustrati schematicamente tutti gli isotopi o specie nucleari, che si trovano in natura, di nove elementi, dal carbonio al silicio. Tre di essi (fluoro, sodio e alluminio) hanno un solo isotopo stabile. I sei isotopi della prima colonna, che hanno un ugual numero di protoni *(pallini in nero)* e di neutroni *(cerchietti in nero)*, sono quelli che più facilmente vengono sintetizzati nelle stelle e pertanto sono i membri più abbondanti delle loro specie. Gli isotopi della seconda colonna hanno un neutrone in più *(cerchietti in colore)* e quelli della terza colonna hanno due neutroni in più. Poiché la loro sintesi è più difficile, gli isotopi che presentano neutroni in più sono meno abbondanti. Tuttavia i nuclei della prima colonna, eccettuato l'azoto 14 (N-14), vengono spesso chiamati nuclei di particelle alfa poiché possono essere considerati costituiti da particelle alfa (nuclei di elio 4, formati da due protoni e due neutroni). I nuclei hanno una carica positiva Z, che corrisponde al numero di protoni. Protoni e neutroni vengono chiamati collettivamente nucleoni.

per nuclei che arrivano da direzioni diverse. I nuclei di raggi cosmici arrivano in modo uniforme da tutte le direzioni. Se si vuole misurare la massa di un nucleo con una precisione superiore all'1 per cento, si deve misurare la lunghezza del suo percorso con una precisione ancora superiore. La caratteristica che permette agli strumenti costruiti a Berkeley e al Cal Tech di identificare isotopi di elementi, mentre analoghi strumenti progettati precedentemente potevano identificare solo elementi, è stata l'aggiunta di rilevatori sensibili di posizione all'ingresso dello spettrometro. Nello strumento progettato al Cal Tech le traiettorie sono determinate da una combinazione di due rilevatori di posizione allo stato solido; lo strumento progettato a Berkeley si basa, per raggiungere lo stesso scopo, su sei camere «a deriva» sature di gas.

Poiché l'ISEE-3 gira intorno al Sole e non intorno alla Terra e rimane fuori dal campo magnetico terrestre, può osservare in continuo raggi cosmici energetici provenienti dalla Galassia. Quando un raggio cosmico si arresta in uno degli spettrometri, lo strumento registra i segnali provenienti da vari rilevatori e li ritrasmette sulla Terra. Gli spettrometri rispondono a una gran varietà di elementi a partire dall'idrogeno fino al nichel che ha 28 protoni. In questi esperimenti sono stati nostri collaboratori Douglas E. Greiner e Harry H. Heckman del Lawrence Berkeley Laboratory e John D. Spalding e Rochus E. Vogt del Cal Tech. Qui di seguito ci occuperemo della misurazione degli isotopi di neo, magnesio e silicio a causa della loro importanza nello studio della nucleosintesi nelle stelle.

Il neo, il magnesio e il silicio, con 10, 12 e 14 protoni rispettivamente, hanno tre isotopi stabili ciascuno. Nella materia del sistema solare gli isotopi dominanti sono neo 20, magnesio 24 e silicio 28, nuclei costituiti da un ugual numero di neutroni e di protoni. Perciò possono essere considerati composti rispettivamente da cinque, sei e sette particelle alfa: nuclei di elio (He-4) costituiti da due protoni e due neutroni. Tali «nuclei di particelle alfa» sono particolarmente stabili e prodotti in abbondanza nei normali processi stellari.

Esperimenti condotti nella metà degli anni settanta da parte di ricercatori dell'Università del New Hampshire, del Goddard Space Flight Center della NASA e dell'Università di Chicago dimostrarono che il neo, il magnesio e il silicio presenti nei raggi cosmici sono simili alla materia del sistema solare poiché anche in essi sono dominanti gli stessi isotopi costituiti da particelle alfa. Questa scoperta fu importante per dimostrare che la maggior parte dei raggi cosmici non ha origine da processi anomali di nucleosintesi. Tuttavia, per ricavare un maggior numero di dati quantitativi sulla possibilità che la nucleosintesi di raggi cosmici e di materia del sistema solare siano state differenti, è necessario misurare le abbondanze relative degli isotopi più pesanti e più rari di neo, magnesio e silicio, ossia di quegli isotopi che contengono uno o due neutroni in più.

Gli spettrometri installati su ISEE-3 avevano innanzi tutto il compito di fornire dati ben precisi su questi isotopi più rari, ricchi di neutroni. Analizzando i risultati, occorre apportare delle correzioni per il fatto che alcuni nuclei osservati sono secondari, ossia frammenti, derivati dalla scissione di nuclei più pesanti di raggi cosmici che sono entrati in collisione con atomi di gas interstellare. La percentuale di nuclei secondari nel campione complessivo può essere stimata osservando l'abbondanza di alcuni nuclei che sono relativamente rari fra i prodotti di nucleosintesi e che pertanto dovrebbero essere quasi totalmente assenti a meno che non si siano formati per frammentazione. Fra questi

L'abbondanza nel sistema solare di elementi dal carbonio al silicio riflette la distribuzione degli elementi nel gas interstellare dal quale il Sole, i pianeti e gli altri corpi del sistema solare si sono condensati cinque miliardi di anni fa. Gli isotopi sono riportati su una scala logaritmica nella quale l'abbondanza del silicio 28 è pari a 10^6. L'abbondanza si basa principalmente su misurazioni della composizione di corpi solidi del sistema solare. Il grafico si basa su dati raccolti da A. G. W. Cameron della Harvard University. I valori servono da standard rispetto ai quali si può misurare la distribuzione degli isotopi nei raggi cosmici. Qualsiasi differenza si possa rilevare tra la materia del sistema solare e quella dei raggi cosmici può essere attribuita alle differenti condizioni stellari in cui sono stati sintetizzati i due tipi di materia. L'abbondanza di neo 22 nel sistema solare è incerta *(area tratteggiata)* perché sia le meteoriti, sia le particelle emesse dal Sole presentano composizioni isotopiche diverse.

Lo spettrometro di massa progettato per l'analisi dei raggi cosmici identifica elementi e isotopi misurando l'energia che i singoli nuclei perdono quando vengono rallentati e arrestati da due rilevatori allo stato solido *(a sinistra)*. Lo strumento dà segnali elettrici proporzionali all'energia persa in ciascun rilevatore. Essenzialmente, lo spettrometro identifica elementi poiché la loro perdita di energia è all'incirca proporzionale a Z^2/v^2, ossia al rapporto fra il quadrato della carica elettrica e il quadrato della velocità. Isotopi dello stesso elemento vengono separati poiché la loro energia cinetica ($E = \frac{1}{2}mv^2$) è proporzionale alla massa. La fotografia al centro illustra la separazione di isotopi ottenuta con lo spettrometro costruito al Cal Tech durante una taratura presso il Lawrence Berkeley Laboratory. La fonte era un fascio di argo 40 (Ar-40) fornito dall'acceleratore Bevalac. Molti nuclei di argo 40 si divisero in nuclei più leggeri in seguito all'urto contro il bersaglio. I vari nuclei passarono attraverso un rilevatore dello spessore di 0,5 millimetri e si fermarono in un secondo rilevatore dello spessore di 1,7 millimetri. Ciascuno dei 10 000 o più nuclei, dall'alluminio (Z = 13) all'argo (Z = 18), appare come un punto in una delle sei ampie bande. Le strisce strette all'interno di ciascuna banda corrispondono a isotopi separati di ogni elemento. Le misurazioni per identificare gli isotopi devono essere circa 10 volte più precise delle misurazioni per identificare gli elementi. A destra è tracciata la risoluzione di cinque isotopi di cloro.

La determinazione delle traiettorie dei raggi cosmici è essenziale operando con uno spettrometro per isotopi nello spazio, dove i raggi cosmici arrivano da tutte le direzioni. L'energia che il nucleo in arrivo perde nel passare attraverso i rilevatori dipende dalla lunghezza del percorso della particella attraverso lo strumento. L'informazione fornita quando un nucleo passa attraverso due rilevatori di posizione (*P_1, P_2*) permette di calcolare la traiettoria e quindi la lunghezza del percorso.

Il rilevatore di posizione nella fotografia a sinistra ha 24 strisce metalliche parallele su una superficie del wafer circolare di silicio e, perpendicolare a queste, un'altra serie di 24 strisce sull'altra superficie. Il rilevatore fornisce così un insieme di coordinate *x-y* per un raggio cosmico in arrivo. Per mezzo di due di questi rilevatori, uno sull'altro, lo spettrometro del California Institute of Technology determina le traiettorie dei nuclei in arrivo. La fotografia a destra mostra uno dei 10 rilevatori di energia che sono impilati sotto i rilevatori di posizione. Questo rilevatore ha due aree sensibili *(in grigio più chiaro)*. L'anello che circonda il disco centrale serve a rilevare i nuclei che entrano o escono attraverso i lati dello strumento. Lo spettrometro del California Institute of Technology è uno dei due spettrometri portati nello spazio dal Third International Sun-Earth Explorer (ISEE-3). Il secondo strumento, costruito all'Università della California di Berkeley, è provvisto di analoghi rilevatori di energia, ma di un sistema diverso per le traiettorie.

nuclei ci sono quelli di litio, berillio e boro (che hanno rispettivamente tre, quattro e cinque protoni) e isotopi quali ossigeno 17 e neo 21. Fatta eccezione per il neo 21, sembra che i tre quarti degli isotopi ricchi di neutroni di neo, magnesio e silicio siano primari, ossia nuclei provenienti da una fonte di raggi cosmici che sono sopravvissuti intatti al viaggio, della durata di 10 milioni di anni, attraverso lo spazio interstellare; solo circa un quarto degli isotopi è secondario, creato per effetto delle collisioni.

Dopo aver apportato questa correzione, noi troviamo che l'abbondanza isotopica nei raggi cosmici differisce in misura notevole da quella nella materia del sistema solare. La differenza più notevole è costituita dalla prevalenza nei raggi cosmici di neo 22. In campioni di materia del sistema solare ci sono, per 100 nuclei di neo 20, dai 7 ai 12 nuclei di neo 22. Nella fonte di raggi cosmici, invece, noi troviamo circa 50 nuclei di neo 22 per 100 nuclei di neo 20. Oltre a ciò, i quattro isotopi ricchi di neutroni del magnesio e del silicio (Mg-25, Mg-26, Si-29 e Si-30) sono più abbondanti del 60 per cento circa nella fonte di raggi cosmici, di quanto non lo siano nei campioni di materia del sistema solare. Benché l'eccedenza di neo 22 nei raggi cosmici sia stata precedentemente riportata dai ricercatori dell'Università del New Hampshire, del Goddard Space Flight Center della NASA e dell'Università di Chicago, questi primi esperimenti non erano stati abbastanza sensibili da rilevare aumenti più piccoli degli isotopi ricchi di neutroni del magnesio e del silicio.

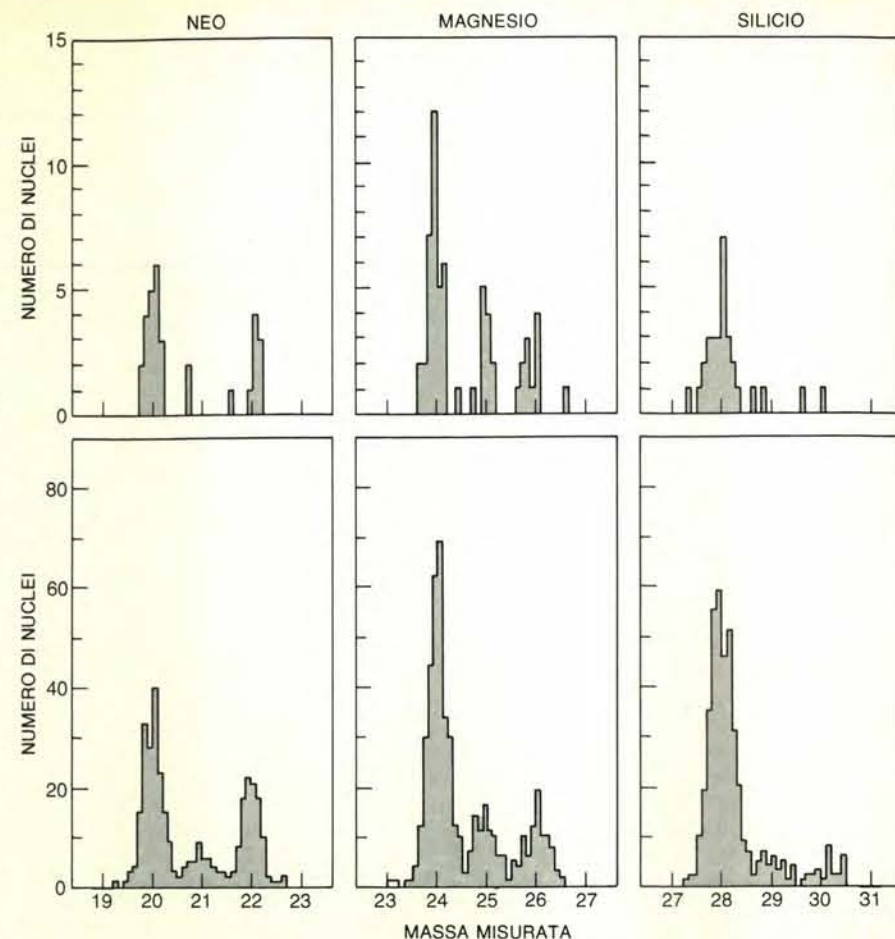
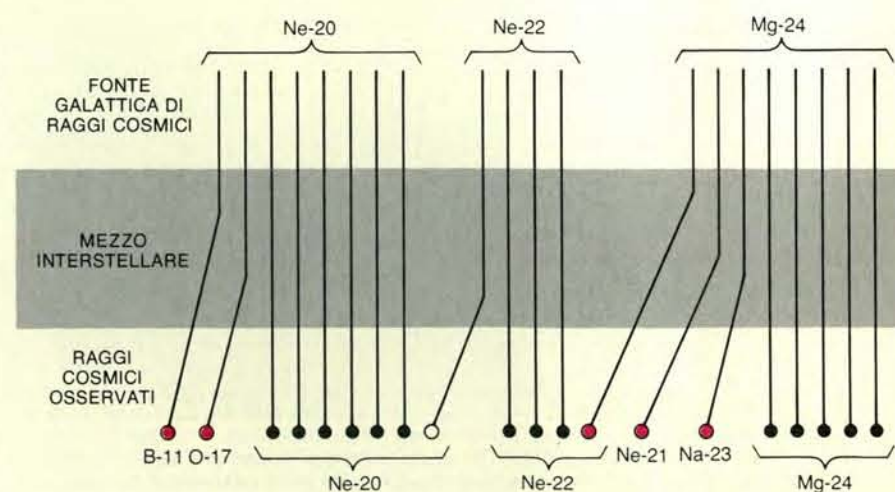
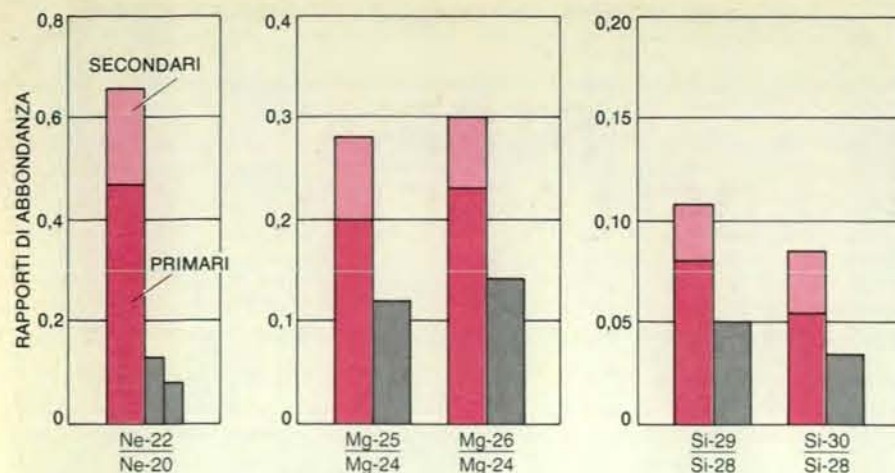
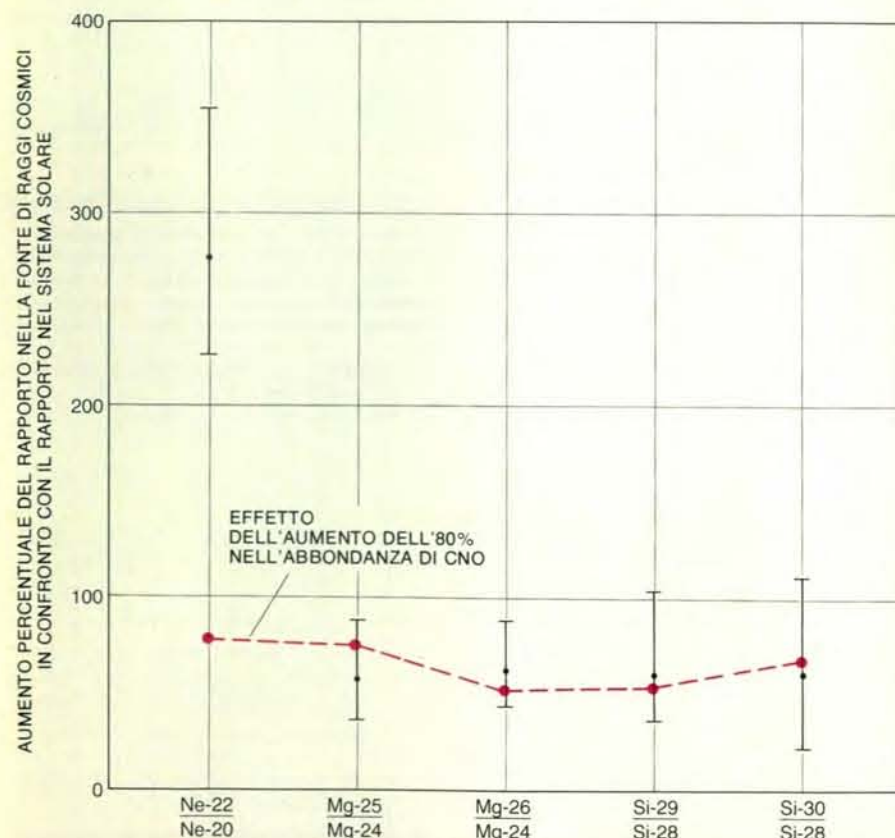
La dimensione delle anomalie riscontrate nei raggi cosmici, che variano dal 60 per cento a diverse centinaia per cento, è impressionante quando si considera che le anomalie isotopiche riportate in anni recenti per alcune meteoriti sono generalmente dell'ordine dell'1 per cento o meno per quanto riguarda il rapporto di abbondanza tra due isotopi, ad esempio tra magnesio 25 e magnesio 24. In altri campioni di sistema solare una eccezione a quella che si può definire uniformità è rappresentata dal rapporto fra neo 22 e neo 20, che varia quasi di un fattore di due. L'interpretazione più ragionevole che si può dare alle grandi anomalie riscontrate nei raggi cosmici è che molti dei nuclei che finiscono come raggi cosmici si siano sintetizzati in condizioni diverse da quelle che portarono alla formazione di nuclei tipici del sistema solare. Per poter capire come possono aver avuto origine tali differenze, dobbiamo considerare alcuni processi che intervengono nella nucleosintesi stellare.

Gli elementi vengono sintetizzati nelle stelle quando nuclei più leggeri si fondono per formare nuclei più pesanti. Quando nuclei più leggeri formano elementi pesanti come il ferro, il processo di fusione trasforma massa in energia e quindi viene chiamato spesso «combustione» dei nuclei più leggeri. Ci sono due fasi differenti nel processo di fusione. La maggior parte dei nuclei pesanti si forma per lenta, quiescente fusione di nuclei sempre più pesanti nelle stelle comuni, un processo che libera energia termonucleare nell'arco di milioni o di miliardi di anni.

Un tipo completamente diverso di nucleosintesi avviene in stelle che superano una certa massa e che hanno esaurito la loro riserva di nuclei capaci di liberare energia mediante il processo di fusione. Privata di una persistente fonte di energia, la stella collassa e poi esplode come supernova. In questa fase, in cui il combustibile è fornito dalla liberazione di energia gravitazionale, viene sintetizzato un ampio spettro di nuclei pesanti, letteralmente in un lampo. La parte maggiore della massa stellare viene espulsa e dispersa nel mezzo interstellare circostante, lasciando dietro di sé un oggetto denso, compatto: una stella di neutroni o, forse, un buco nero.

Poiché i due processi di nucleosintesi, quello quiescente e quello esplosivo, sono caratterizzati da temperature e da tempi completamente diversi, essi danno origine a nuclei del tutto differenti. La maggior parte dei nuclei con un ugual numero di protoni e di neutroni è sintetizzata attraverso una successione di processi di combustione quiescente, che inizia con la fusione di nuclei di idrogeno (protoni) in nuclei di elio comune, He-4. In questi processi la metà dei protoni viene trasformata in neutroni per decadimento radioattivo in cui la carica positiva di un protone viene emessa sotto forma di un positone, o elettrone positivo. I nuclei di idrogeno si combinano anche con qualsiasi traccia di carbonio, azoto, o ossigeno (chiamati collettivamente nuclei di CNO) che possono essere preesistiti nella materia dalla quale si è formata la stella. Perché siano presenti, i nuclei di CNO devono essere stati creati in una generazione precedente di nucleosintesi stellare e iniettati nel mezzo interstellare che diede origine alla stella. Nella fase di combustione dell'idrogeno, quasi tutti i nuclei di carbonio e di ossigeno sono trasformati in azoto ordinario, N-14, mediante una serie di reazioni nucleari nota come ciclo di CNO.

Il Sole trasforma continuamente l'idrogeno in elio e continuerà a farlo per altri cinque milioni di anni. Una stella di massa maggiore brucerà il suo idrogeno più rapidamente poiché la sua temperatura interna e la sua densità sono più elevate. Al termine della fase di fusione dell'idrogeno, la parte centrale di una stella massiva si contrae, raggiungendo una temperatura sufficientemente elevata da bruciare elio, per dare principalmente carbonio 12 e ossigeno 16 mediante la fusione di tre o quattro nuclei di elio, rispettivamente. In questa fase del ciclo vitale della stella, l'elio si combina anche con l'azoto 14 secondo una serie di reazioni di fusione e successivi decadimenti radioattivi il cui prodotto finale è neo 22, un nucleo che ha due neutroni in più rispetto ai protoni.

Il numero totale dei neutroni in più in una stella è significativo poiché regola la produzione di altri nuclei ricchi di neutroni sintetizzati in entrambe le fasi della vita di una stella: quella quiescente e quella finale esplosiva. Il neo 22 deriva dall'azoto 14, il quale a sua volta deriva dalla provvista originaria di nuclei di CNO nella stella, così che l'abbondanza di neo 22 è proporzionale

La distribuzione degli isotopi nei raggi cosmici, come è stata determinata nello spazio dallo spettrometro del Cal Tech *(in alto)* e da quello di Berkeley *(in basso)* mostra che neo, magnesio e silicio sono tutti rappresentati dai loro tre isotopi stabili; sono dominanti gli isotopi costituiti da particelle alfa: neo 20, magnesio 24 e silicio 28. I 120 eventi nel campione del Cal Tech furono raccolti nei primi tre mesi e mezzo di osservazione, prima che lo spettrometro fosse parzialmente fuori uso. I 1226 eventi nel campione di Berkeley rappresentano i dati accumulati in un periodo di due anni dopo il quale si guastarono i sensori di traiettoria. Ai dati forniti dallo spettrometro di Berkeley viene quindi dato un peso maggiore quando i risultati sono confrontati. Gli esperimenti sono stati condotti dagli autori e dai loro colleghi, compresi Douglas E. Greiner e Harry H. Heckman del Lawrence Berkeley Laboratory e John D. Spalding e Rochus E. Vogt del Cal Tech.

La frammentazione dei raggi cosmici per effetto delle collisioni nel mezzo interstellare altera la composizione dei nuclei originariamente accelerati alla fonte di raggi cosmici. Tali frammenti di collisione sono detti secondari. Lo schema illustra alcuni nuclei secondari *(in colore)* che si possono formare quando i nuclei primari di tre isotopi (neo 20, neo 22 e magnesio 24) entrano in collisione con atomi di gas interstellare. Isotopi, quali boro 11, ossigeno 17 e neo 21, sono relativamente rari tra i prodotti della nucleosintesi stellare e possono essere considerati come secondari. La loro abbondanza dà la chiave per calcolare la frammentazione. A eccezione del neo 21, i tre quarti dei nuclei ricchi di neutroni di neo, magnesio e silicio sono di origine primaria.

L'abbondanza di isotopi ricchi di neutroni nei raggi cosmici è riportata in colore sotto forma di rapporti in cui al denominatore appaiono gli isotopi costituiti da particelle alfa del neo, del magnesio e del silicio. L'abbondanza corrispondente nella materia del sistema solare è rappresentata dalle barre in grigio. Nei rapporti che riguardano i raggi cosmici, è stata fatta una distinzione fra isotopi primari e secondari. Poiché il neo del sistema solare presenta numerosi rapporti di abbondanza isotopica, la gamma di variazione è indicata da due barre. La misurazione di raggi cosmici è la media ponderata dei risultati ottenuti a Berkeley e al California Institute of Technology. Esse mostrano che la fonte ha un eccesso di isotopi di questi elementi ricchi di neutroni.

L'eccesso di isotopi ricchi di neutroni nella materia che costituisce la fonte dei raggi cosmici riflette forse i processi di nucleosintesi in stelle la cui composizione iniziale differiva dalla composizione della generazione di stelle che fornì la materia al sistema solare. Secondo S. E. Woosley dell'Università della California a Santa Cruz e T. A. Weaver del Lawrence Livermore National Laboratory l'eccesso di isotopi ricchi di neutroni può essere spiegato se le stelle responsabili dei nuclei nei raggi cosmici avevano inizialmente un'abbondanza di nuclei di carbonio, azoto e ossigeno (chiamati collettivamente nuclei di CNO) dell'80 per cento superiore a quella delle stelle responsabili dei nuclei del sistema solare. La linea tratteggiata in colore mostra l'aumento di isotopi, ricchi di neutroni, di neo, magnesio e silicio che dovrebbe esserci nei raggi cosmici provenienti da tali stelle. La linea tratteggiata attraversa quattro delle cinque barre che descrivono l'aumento osservato degli isotopi ricchi di neutroni dei raggi cosmici in confronto a quello degli stessi isotopi del sistema solare. Tuttavia il rapporto tra neo 22 ricco di neutroni e neo 20 è anomalo per più del 200 per cento anche se il valore massimo per il sistema solare è assunto come base per il confronto. L'estensione verticale delle barre indica le incertezze sperimentali nelle osservazioni dei raggi cosmici.

alla originaria abbondanza di CNO. Ad esempio, il 2 per cento della materia del sistema solare è rappresentato da nuclei di CNO. Pertanto una stella, che in origine conteneva un'analoga abbondanza di CNO, dopo il completamento della fusione dell'idrogeno e dell'elio, terminerà con un contenuto del 2 per cento di neo 22 nella sua parte centrale. Benché ogni nucleo di neo 22 abbia un eccesso di neutroni pari a circa il 10 per cento, il grosso della materia sarà costituito da carbonio 12 e ossigeno 16, nuclei che non hanno neutroni extra, così che l'eccesso complessivo di neutroni sarà solo dello 0,2 per cento, che corrisponde a due neutroni in più, rispetto al numero dei protoni, ogni 1000 nucleoni.

Quando nella stella l'elio si avvicina all'esaurimento, la stella si contrae di nuovo e diventa più calda, rendendo così possibile la fusione dei nuclei di carbonio e di ossigeno con formazione di nuclei quali neo 20, magnesio 24 e silicio 28, che hanno tutti un ugual numero di protoni e di neutroni e che sono gli isotopi dominanti di quegli elementi. Tali nuclei di particelle alfa sono i prodotti principali della fase quiescente di nucleosintesi. In questa fase alcuni atomi di neo 22 vengono trasformati anche in magnesio 25, con liberazione di neutroni, che possono essere catturati da altri nuclei creando così isotopi supplementari ricchi di neutroni.

Quando la parte centrale di una stella massiccia si svuota di nuclei leggeri capaci di promuovere la combustione quiescente, la stella diventa instabile e ne segue un'esplosione di supernova, che trasmette un'onda d'urto dalla profondità più interna della stella alla materia sovrastante e per pochi decimi di secondo ne innalza la temperatura a diversi miliardi di gradi. In questo breve periodo si sintetizzano molti nuclei supplementari, tra i quali magnesio 25, magnesio 26, silicio 29 e silicio 30. La quantità di questi isotopi ricchi di neutroni che si sono creati è determinata dall'abbondanza, precedente alla esplosione, di neo 22, che è una fonte di neutroni extra. Ne risulta che l'eccesso di neutroni assicurato in origine dall'abbondanza di nuclei di CNO determina la proporzione di nuclei ricchi di neutroni che si sono sintetizzati nel periodo quiescente e in quello esplosivo della vita di una stella.

Alcuni anni fa, W. David Arnett, attualmente all'Università di Chicago, propose la teoria secondo la quale la Galassia mostra un costante aumento del numero di nuclei ricchi di neutroni come risultato dell'effetto cumulativo di nucleosintesi in successive generazioni di stelle. Poiché i raggi cosmici, che arrivano in vicinanza della Terra oggi, sono presumibilmente molto più giovani del materiale che forma il sistema solare, essi dovrebbero riflettere l'aumento di nuclei ricchi di neutroni. Tuttavia, recenti modelli di evoluzione galattica suggeriscono che l'atteso arricchimento di isotopi ricchi di neutroni nella Galassia nel suo complesso, nel corso dei cinque miliardi di anni trascorsi, possa non essere sufficiente a spiegare le nostre osservazioni sui raggi cosmici, particolarmente nel caso del neo 22.

Più recentemente S. E. Woosley dell'Università della California a Santa Cruz e T. A. Weaver del Lawrence Livermore National Laboratory hanno proposto la teoria secondo la quale l'eccesso osservato di isotopi ricchi di neutroni nei raggi cosmici galattici può essere spiegato se vi sono nella nostra galassia regioni in cui l'abbondanza di nuclei di CNO è notevolmente maggiore di quanto lo sia nella materia del sistema solare. I loro calcoli suggeriscono che la materia espulsa dalle supernove, che inizialmente aveva un'abbondanza di CNO maggiore dell'80 per cento rispetto a quella della materia del sistema solare, si sarebbe arricchita di circa l'80 per cento di neo 22 e di un valore variabile dal 50 al 70 per cento di ciascuno degli isotopi ricchi di neutroni del magnesio e del silicio. Questo modello è suffragato dall'arricchimento pressoché uguale che noi troviamo per quattro dei cinque isotopi ricchi di neutroni. Pertanto l'eccedenza del 60 per cento di magnesio 25, magnesio 26, silicio 29 e silicio 30 può essere adeguatamente intesa come il risultato di un aumento dell'80 per cento della abbondanza di nuclei di CNO nelle regioni in cui hanno origine i raggi cosmici.

Tuttavia, l'eccedenza di neo 22 che noi troviamo supera il 200 per cento ed è pertanto in netto contrasto con l'aumento che ci si aspettava dell'80 per cento, in conseguenza dell'aumento dell'abbondanza di nuclei di CNO nella Galassia, aumento che è pari appunto all'80 per cento. L'eccedenza di neo 22 osservata può derivare da qualche aspetto di nucleosintesi trascurato nei modelli semplici di evoluzione chimica della Galassia, ma può anche derivare da una mancanza di cognizioni esatte sulla abbondanza degli isotopi del neo nella materia del sistema solare. Ad esempio, numerose composizioni ben definite di isotopi del neo sono state misurate in meteoriti e nel Sole.

Essa potrebbe derivare anche da una classe limitata di oggetti stellari che sintetizzano neo 22 con un'abbondanza di molto superiore a quella trovata nel sistema solare. Alcuni ricercatori hanno proposto modelli in cui potrebbe essere sintetizzato un largo eccesso di neo 22, mentre viene mantenuta una miscela di altri nuclei molto simile a quella del sistema solare. Non è ancora evidente, tuttavia, se tali fonti possano spiegare le altre anomalie isotopiche che abbiamo misurato in magnesio e silicio.

Una proposta più radicale è stata avanzata recentemente da K. A. Olive e David N. Schramm dell'Università di Chicago. Secondo la loro teoria, i raggi cosmici riflettono accuratamente la composizione del mezzo interstellare della Galassia in generale ed è il sistema solare che è atipico. Essi rivolgono l'attenzione a studi recenti compiuti su meteoriti i quali indicano che il sistema solare, quando si formò, conteneva un isotopo radioattivo a vita breve dell'alluminio, Al-26. L'esistenza dell'alluminio 26 è desunta dalla presenza nelle meteoriti di un suo prodotto di decadimento, il magnesio 26, in associazione con l'alluminio 27, stabile. L'alluminio 26 ha un periodo di dimezzamento di solo un milione di anni circa, così che non potrebbe essere stato

L'abbondanza relativa di isotopi nella Galassia è determinata dalla formazione, evoluzione e distruzione esplosiva di stelle massive. In una regione della Galassia in cui la composizione del gas interstellare *(a)* è simile a quella della materia nel sistema solare una nube di gas collassa per effetto della sua stessa gravità dando origine a una nuova stella *(b)*. All'interno della stella la fusione termonucleare trasforma parte dell'idrogeno e dell'elio originari in nuclei di particelle alfa, come carbonio 12 e ossigeno 16 *(c)*. Contemporaneamente i nuclei di CNO, in origine presenti nel combustibile stellare, sono trasformati in nuclei più pesanti ricchi di neutroni come neo 22 e magnesio 25. Quando viene esaurito tutto il combustibile nucleare all'interno della stella, questa esplode come supernova *(d)*. L'onda d'urto generata dall'esplosione provoca la sintesi di altri nuclei pesanti ed espelle la maggior parte dei prodotti di nucleosintesi nel gas interstellare. Questa sequenza di eventi, ripetuta in successive generazioni di stelle, fa sì che il gas interstellare si arricchisca sempre più in carbonio, azoto e ossigeno e in nuclei pesanti con un eccesso di neutroni. Alcuni nuclei nel gas sono accelerati alla velocità dei raggi cosmici, forse a causa delle onde d'urto provenienti da supernove *(e)*. L'accelerazione dei raggi cosmici potrebbe anche avvenire direttamente, quando la supernova espelle materia nello spazio interstellare *(d)*.

presente quando il sistema solare si condensò dal gas interstellare, a meno che non sia stato creato di recente. La spiegazione più probabile è che un'esplosione di supernova abbia immesso l'alluminio 26 proprio prima della formazione del sistema solare. In quel caso, altri elementi, espulsi anch'essi dalla supernova, potrebbero aver conferito al sistema solare una composizione notevolmente diversa dalla composizione del mezzo interstellare medio.

Olive e Schramm propongono un modello specifico in cui la concentrazione di nuclei ricchi di neutroni, tipici del mezzo interstellare al tempo in cui si formò il sistema solare, venne diluita per effetto dell'immissione di un eccesso di nuclei di particelle alfa derivate da una o più fonti di supernova. Tale abbondanza modificata sarebbe pertanto in disaccordo con l'abbondanza misurata nei raggi cosmici, facendo apparire i raggi cosmici ricchi di neutroni conformi, da un punto di vista qualitativo, con le nostre osservazioni. Se Olive e Schramm hanno ragione, la reale importanza delle osservazioni di raggi cosmici può trovarsi nella luce che essi gettano sugli eventi associati con la formazione del sistema solare.

Come mostrano questi vari modelli e proposte, la miscela di elementi e di isotopi in un campione di materia dipende dalla natura della nucleosintesi nelle stelle, dalla storia della formazione di una stella e dall'evoluzione della Galassia in generale. La comprensione teorica di tali processi sta avanzando e approfondendosi per effetto di osservazioni minuziose di campioni di materia che i raggi cosmici portano da altre regioni della Galassia. La conoscenza finora acquisita è ancora molto limitata. L'abbondanza isotopica di tre elementi soltanto, neo, magnesio e silicio, è stata determinata accuratamente nella fonte di raggi cosmici, e in ciascun caso si è trovato che differisce dall'abbondanza corrispondente nei campioni di sistema solare. Forse le anomalie finora osservate potranno diventare la regola, e non l'eccezione.

Molti altri elementi, come ferro e nichel, hanno numerosi isotopi stabili la cui abbondanza servirà come da sonda per la nucleosintesi. Si sa già che l'isotopo più abbondante del ferro, sia nel sistema solare, sia nei raggi cosmici, è Fe-56. Tuttavia non ci sono tuttora precisi dati quantitativi sull'abbondanza nei raggi cosmici di due altri isotopi del ferro, Fe-54 e Fe-58, che sono sintetizzati in rapporti diversi in differenti condizioni stellari. Gli strumenti che verranno installati sui futuri veicoli spaziali e che saranno in grado di misurare con precisione questi e altri isotopi, aiuteranno a risolvere questi enigmi.

He-4 He-4 He-4 → FUSIONE → C-12 → FUSIONE (He-4) → O-16 → FUSIONE (He-4) → Ne-20
RAGGIO GAMMA RAGGIO GAMMA RAGGIO GAMMA

La fusione quiescente dei nuclei di elementi leggeri crea nuclei di massa maggiore. Nel primo stadio (non illustrato), nuclei di idrogeno si fondono e formano elio (particelle alfa) ed essenzialmente tutti i nuclei di CNO che potevano essere presenti nel combustibile originario sono trasformati in azoto 14. Nella sequenza a sinistra è illustrato il secondo stadio, in cui nuclei di elio si fondono e formano nuclei di particelle alfa, quali carbonio 12, ossigeno 16 e neo 20. La sequenza a destra illustra i processi di fusione e di decadimento radioattivo che trasformano i nuclei di azoto 14 (che sono stati sintetizzati prima da nuclei di carbonio) in nuclei di neo 22. Pertanto l'abbondanza finale di neo 22, che ha due neutroni in più rispetto ai protoni, dipende dall'abbondanza iniziale di carbonio, azoto e ossigeno nel combustibile della stella.

La nucleosintesi esplosiva avviene durante l'esplosione di supernova di una stella quando è terminata la fase di combustione quiescente. L'onda d'urto esplosiva porta la temperatura della materia stellare fino a diversi miliardi di gradi centigradi per pochi decimi di secondo, dando inizio a una complessa serie di reazioni nucleari, alcune qui illustrate. La fusione di neo 22 ed elio 4 *(a sinistra)* è particolarmente importante poiché i neutroni liberati possono essere catturati nella sequenza di reazioni *(a destra)* che danno isotopi, ricchi di neutroni, di magnesio e silicio. L'abbondanza di questi isotopi è pertanto determinata dalla quantità di neo 22, che dipende dall'abbondanza originaria di CNO nella stella. Anche se una analoga sintesi di isotopi ricchi di neutroni avviene in una fase di combustione quiescente, le velocità di reazione sono molto inferiori.

Organismi sessili e movimenti dell'acqua

Un gran numero di organismi marini che vivono fissi al fondo in prossimità delle coste mostrano una notevole varietà di adattamenti che permettono loro di affrontare con successo onde e maree

di M. A. R. Koehl

Chiunque cammini lungo una costa rocciosa non riparata potrà notare, durante la bassa marea, che attaccati al fondo vivono una notevole quantità e varietà di piante e di animali destinati a essere battuti dalle onde e che corrono il rischio di venire strappati dal substrato dai movimenti dell'acqua; d'altra parte, dipendono tutti dall'acqua per quanto riguarda il nutrimento, l'eliminazione dei rifiuti e la diffusione della loro progenie verso nuove zone. Come se la cavano?

Ciò che si impara esaminando da vicino gli animali e le piante sessili, come è capitato a me di fare lungo le coste del Pacifico dell'America Settentrionale e Meridionale e lungo le scogliere coralline dei Caraibi, è che gli organismi attuano una serie di compromessi per ottenere il massimo dai movimenti dell'acqua riducendo al minimo gli effetti dannosi. Organismi che vivono in mezzo a un flusso lento hanno spesso caratteristiche atte a compensare le carenze di trasporto da parte dell'acqua; per esempio sono in grado di realizzare correnti nutritive loro proprie e possono tollerare basse concentrazioni di ossigeno. Organismi che vivono in mezzo a flussi veloci hanno caratteristiche rivolte ad aumentare le loro capacità di adesione al substrato e di utilizzazione dell'acqua che li oltrepassa rapidamente. Gli adattamenti che consentono agli organismi sessili di restare fissi in un luogo sono i più svariati che si possano immaginare.

Il ricercatore che voglia studiare il modo in cui un organismo sessile riesce a cavarsela, dal punto di vista meccanico, con i movimenti dell'acqua si trova di fronte a una serie di domande. In che modo il flusso dell'acqua e le forze esercitate da tale passaggio su un organismo vengono influenzati dalla struttura di questo? Che relazione esiste fra la forma di un organismo e le sollecitazioni che vengono esercitate sui suoi tessuti dalle forze di flusso? In che maniera il materiale strutturale di un organismo determina le modalità con le quali lo stesso organismo si deforma o si frattura in risposta alle sollecitazioni prodotte dalle forze di flusso?

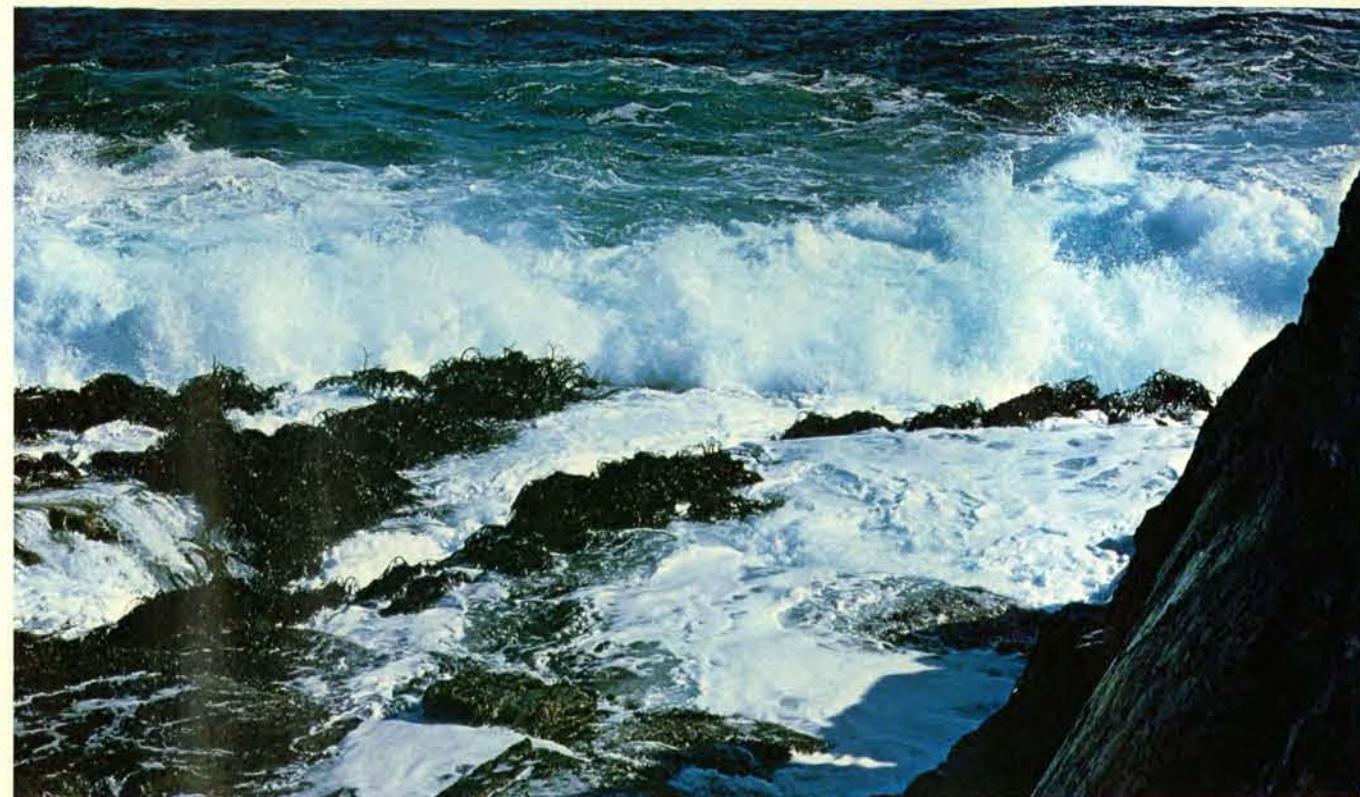
Gli organismi marini sessili vivono in un gran numero di tipi diversi di flusso. Esiste una distinzione tra gli ambienti intercotidali e quelli subcotidali, cioè tra gli ambienti compresi tra la linea di bassa marea e quella di alta marea e gli ambienti oltre il limite della bassa marea. Una pianta o un animale che vive in ambiente intercotidale, in un luogo dove le onde si frangono su una costa non riparata, è soggetto a un flusso veloce e turbolento. Ho misurato il flusso dell'acqua su questi organismi e ho trovato che la velocità più elevata si ha quando un frangente si abbatte sopra la riva e nel movimento di risacca.

Per gli organismi sessili che vivono nella parte alta della zona subcotidale di una costa spazzata dalle onde, l'andirivieni del flusso e del riflusso cui sono sottoposti è più lento. (Si ricordi che la posizione di un organismo rispetto ai frangenti, e di conseguenza rispetto al tipo di flusso cui è sottoposto, varia con il salire e l'abbassarsi della marea.) Gli organismi che vivono a profondità che superano la metà della lunghezza d'onda, cioè della distanza tra le creste di due onde successive, non «sentono» le onde, ma, come quelli che vivono nelle baie e negli stretti, sono esposti a correnti costanti, oppure alle correnti di marea che si invertono periodicamente.

Quelli che ho descritto finora potrebbero essere definiti microhabitat di flusso. Ogni microhabitat è costituito da microhabitat limitati, nei quali il flusso è diverso. Per studiare un microhabitat bisogna prendere in considerazione un aspetto basilare dello scorrimento di fluidi sopra una superficie solida: il flusso è più lento in prossimità della superficie stessa. Questo strato di fluido in lento movimento è detto strato limite; quanto più lento è il flusso o quanto più lontano è il punto di osservazione sulla superficie di un oggetto rispetto al punto in cui questo è investito dal flusso, tanto più spesso risulta lo strato limite. Organismi estremamente piccoli (larve che si sono stabilite in un luogo da poco tempo, piante e animali unicellulari che dimorano sul fondo) e piatti organismi incrostanti (alghe crostose e alcuni briozoi) sono gli abitatori dello strato limite. Il microhabitat di flusso di un piccolo organismo può risultare assai diverso dal flusso principale che interessa l'area dove esso vive. Inoltre, l'ambiente marino di solito non è piatto e levigato; gli scogli, le fessure e le spaccature nelle rocce, e persino altri organismi che si possono trovare intorno a una grossa pianta o a un grosso animale possono alterare notevolmente il flusso che li investe.

Gli anemoni di mare del genere *Anthopleura,* comuni sulle coste rocciose che si estendono dalla California alla British Columbia, offrono un buon esempio di animali che vivono in singolari microhabitat di flusso. *A. xanthogrammica* tappezza il fondo di canali battuti da frangenti intercotidali, in punti particolarmente esposti. Fissi al fondo della zona intercotidale più bassa, questi antozoi sopravvivono nutrendosi di molluschi e ricci di mare che le onde hanno staccato dal substrato e portato fino a loro.

Ho misurato sia i flussi della corrente principale di questi canali agitati sia i flussi della zona in cui si trovano gli anemoni. Nella corrente principale il flusso può raggiungere i cinque metri al secondo, nelle giornate calme; a livello dei tentacoli degli anemoni ha valori molto più bassi. Gli individui di *A. xanthogrammica* hanno un diametro che può raggiungere i venti centimetri, ma in un canale non riparato assumono l'aspetto di una frittella, raggiungendo un'altezza media, al di sopra del substrato roccioso, di 2,5 centimetri. In altri termini, evitano il flusso veloce appiattendosi.

Si può trovare *A. xanthogrammica* anche in zone più protette, dove la violen-

Le fasce di alghe marine che vivono lungo le coste rocciose del Cile centrale e meridionale sono esposte alla violenza del mare che esercita forze notevoli sui tessuti delle alghe. Le forze agiscono in una direzione al frangersi dell'onda, e nella direzione opposta durante la risacca. L'alga qui fotografata è *Lessonia nigrescens,* che risponde alle forze esercitate su di essa piegandosi nella direzione del flusso. Insieme a *Lessonia* si trova in queste fasce un'altra alga bruna, *Durvillea antarctica,* che si lascia tirare dalla corrente in modo da assumere una posizione più o meno parallela al fondo marino. La sollecitazione massima che subisce lo stipite di un'alga che si lascia tirare orizzontalmente dal flusso, come *Durvillea,* è notevolmente inferiore a quella che subisce lo stipite di un'alga che viene soltanto piegata, come *Lessonia.*

I grandi anemoni di mare della specie *Anthopleura xanthogrammica* tappezzano le rocce non riparate che costituiscono il fondo di bracci di mare mossi. Questa fotografia è stata scattata nello stretto di Tatoosh Island, al largo dell'estremità nordoccidentale dello stato di Washington. In un braccio di mare mosso questi antozoi, che sono sessili, lottano contro la violenza dell'acqua appiattendosi sul fondo, dove il flusso è meno forte; in luoghi più protetti, dove gran parte dell'energia dell'acqua viene assorbita dagli scogli, assumono una forma più alta.

za delle onde è attutita da rocce al largo. In un habitat di questo tipo gli anemoni hanno un'altezza media di sette centimetri, e riescono così a compensare abbondantemente la minor velocità della corrente principale nelle zone riparate.

Gli anemoni *A. xanthogrammica* delle coste più riparate spesso sono affiancati da anemoni più piccoli della specie *Anthopleura elegantissima,* che vivono in gruppi numerosi e che si riproducono per scissione longitudinale ricoprendo in tal modo le rocce di cloni, come un tappeto. Ho trovato che la velocità dell'acqua che passava tra questi anemoni piccoli e strettamente raggruppati era solamente un decimo della velocità dell'acqua che investiva un grosso anemone solitario che si trovava in mezzo a loro. Il contrasto illustra quanto diversi possano essere i microhabitat di organismi che pur vivono gli uni accanto agli altri.

Molti animali coloniali marini, come coralli e idrozoi coloniali, secernono uno scheletro che funge da supporto per gli individui della colonia (i polipi). Sono gli animali stessi i realizzatori del substrato sul quale vivono e, pertanto, i modificatori dei propri habitat di flusso. Per visualizzare il flusso dell'acqua intorno a rami di corallo e anche a modelli di corallo stesso, John Chamberlain del Brooklyn College della City University di New York e Richard Graus della Cleveland State University fecero degli esperimenti mettendo un colorante nell'acqua. Come ci si poteva aspettare, videro che il diametro, la spaziatura e la disposizione dei rami influenzavano la velocità e la direzione dei movimenti dell'acqua attraverso la colonia. Il fatto più sorprendente fu la scoperta che configurazioni molto diverse delle colonie potevano dare origine allo stesso tipo di flusso sui polipi. I polipi dei rami di gorgonie e di antipatidi vivono in scheletri flessibili che si curvano in differenti direzioni a seconda dell'onda che li investe. Dato che lo scheletro si muove insieme all'acqua, il flusso, rispetto a un singolo polipo, può essere lento anche se la colonia si trova in un ambiente molto mosso.

Le forze esercitate dal flusso su un organismo marino sessile sono qui illustrate schematicamente. A un aumento di velocità lungo una linea di flusso (il percorso descritto da una particella di fluido) corrisponde una diminuzione della pressione, e viceversa. Nei punti di sezione più ampia di un corpo *(a)*, la velocità del flusso aumenta e la pressione esercitata sulla superficie del corpo diminuisce. Oltre i punti di sezione maggiore il flusso rallenta e la pressione aumenta. La viscosità provoca una perdita di quantità di moto da parte delle particelle di fluido dello strato limite (le particelle più vicine al corpo) mentre queste passano attorno al corpo. Quando il flusso intorno al corpo è sufficientemente veloce e il corpo è sufficientemente grande, il rallentamento dovuto alla viscosità giunge ad arrestare lo spostamento nel senso della corrente dell'acqua dello strato limite, che, pertanto, può essere spinto controcorrente dall'aumento di pressione dovuto al rallentamento dell'acqua che lo circonda. Dietro il corpo si separa dalla corrente principale una scia di turbolenza, facendo sì che la pressione in questa zona sia inferiore a quella esercitata sulla parte frontale. Questa pressione netta nella direzione del flusso viene definita resistenza di forma. Se il fluido si muove con sufficiente velocità rispetto al corpo *(b)* lo strato limite diviene turbolento e si ha un trasferimento di quantità di moto dal flusso di corrente principale a esso; questo provoca una scia minore e una minore resistenza di forma. A mano a mano che il fluido si muove più velocemente sulla parte apicale o su un lato di un corpo *(c)* la bassa pressione ivi esercitata tende a risucchiare il corpo in alto o in direzione laterale. Questa forza viene detta forza di sollevamento.

La forma, le dimensioni e i tipi di tessuti di un organismo sessile influenzano le forze meccaniche esercitate su di esso dall'acqua che lo investe. Le forze, che tendono a spostare un organismo nel senso della corrente, sono dovute alla viscosità dell'acqua, cioè alla tendenza delle molecole d'acqua a resistere allo scivolamento delle une sulle altre. L'acqua bagna le superfici solide, comprese le superfici degli organismi, e uno strato limite d'acqua che circonda un organismo è soggetto a forze di taglio al passaggio del flusso. Dato che l'acqua è viscosa, e oppone resistenza, l'organismo è sottoposto a una forza dovuta all'attrito superficiale. L'entità di questa forza è proporzionale alla velocità del flusso e alle dimensioni dell'organismo. La forza che si esercita su piccoli organismi, in acque in cui il flusso sia lento, è dovuta in gran parte all'attrito superficiale. Organismi più grandi, che si trovino in acque con un flusso veloce, sono soggetti a un'ulteriore forza, dipendente dalla forma (detta «resistenza di forma»), che generalmente è maggiore di quella dovuta all'attrito superficiale. A valle dell'organismo si forma una scia turbolenta, che tende a trascinare l'individuo con sé. Ogni caratteristica dell'organismo che riduca la scia riduce anche la resistenza di forma. L'entità di quest'ultima è proporzionale alla sezione maestra del corpo e al quadrato della velocità, cosicché un lieve aumento delle dimensioni o della velocità porta a un considerevole aumento della resistenza di forma.

Per studiare il modo in cui la forma, la struttura superficiale e la flessibilità di un organismo possono influenzare le forze di flusso si possono seguire numerose vie. Si possono comparare queste forze e gli andamenti del flusso intorno a organismi simili che differiscono per alcune caratteristiche. Si possono anche costruire modelli degli organismi, modificando determinate caratteristiche per valutarne gli effetti sulla resistenza. Gli organismi sessili e i modelli vengono attaccati a trasduttori di forze, sul litorale o in una vasca a flusso, in modo che la velocità di flusso sia controllata. L'andamento del flusso intorno al corpo viene reso evidente mediante particelle di un colorante o di un marcatore, oppure viene analizzato per mezzo di piccole sonde elettroniche.

Se un organismo di grosse dimensioni ha la maggior parte della superficie del proprio corpo disposta parallelamente

Gli effetti risultanti dalle forze che il flusso esercita su un organismo marino sessile sono lo stiramento longitudinale, lo stiramento trasversale, l'incurvamento e la torsione. Le frecce in nero indicano le forze che agiscono sull'organismo, le frecce in colore le deformazioni dei tessuti.

alla direzione del flusso darà origine a una scia più piccola di quella che formerebbe se fosse disposto ad angolo retto, e la resistenza di forma risulterà minore. Questo semplice modo di ridurre la resistenza di forma è ben esemplificato dalle forze che vengono esercitate su due grandi anemoni di mare: il tozzo *A. xanthogrammica,* che ho già menzionato, e il lungo e soffice *Metridium senile,* che possiede una corona di piccoli tentacoli larga e solcata, che serve da filtro per la cattura dello zooplancton. *M. senile* vive in una zona subcoticale, dove le lente correnti di marea piegano la sua corona ad angolo retto rispetto al flusso della corrente. Per contro, *A. xanthogrammica* sta in una posizione in cui la maggior parte della superficie del suo corpo è parallela alla direzione del flusso. A una certa velocità tanto le dimensioni della scia quanto l'entità della resistenza di forma sono maggiori per *M. senile* che per *A. xanthogrammica,* a parità di diametro della corona di tentacoli.

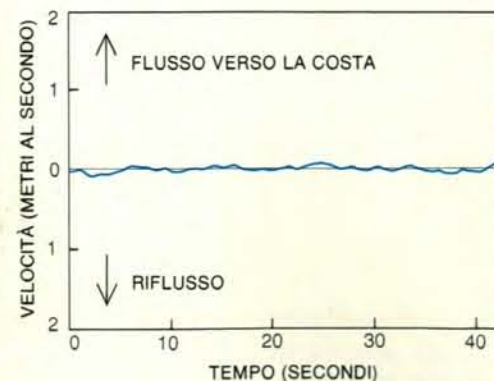
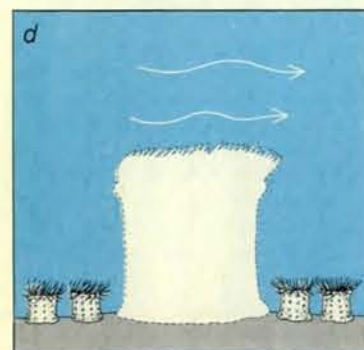
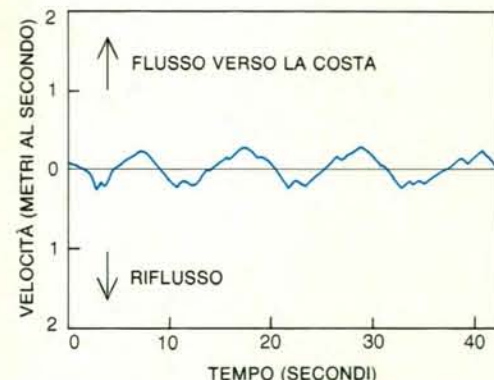
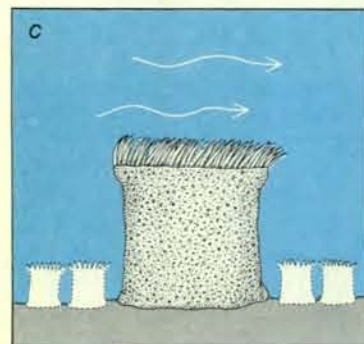
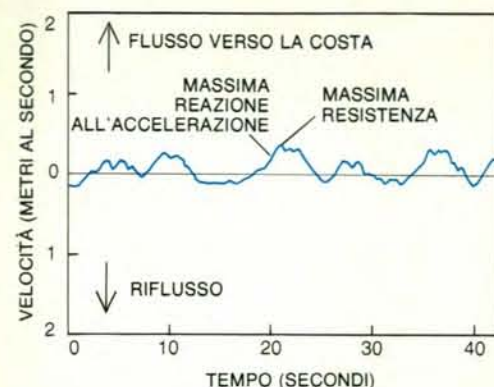
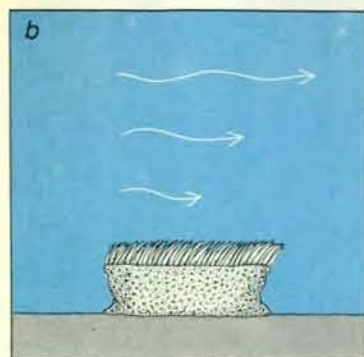
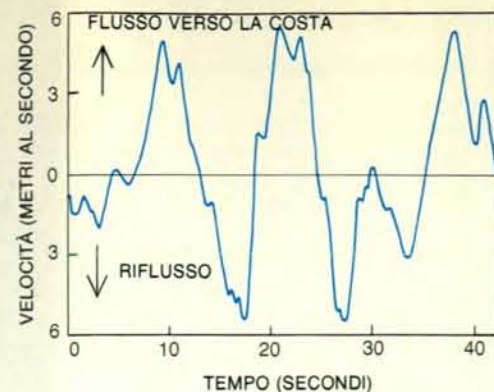
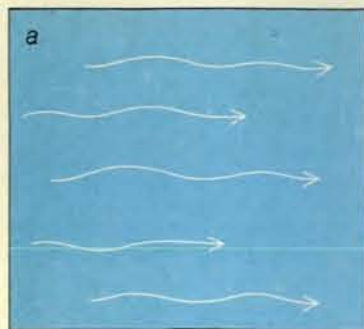
Le strutture ramificate e i tentacoli di molti animali marini e di molte piante flessibili vengono spinti a raggrupparsi quando l'organismo è investito dal flusso d'acqua. Tale movimento mette l'organismo in una posizione che meglio asseconda il flusso, riducendo la resistenza e la scia. *M. senile* ne è un esempio: a mano a mano che la velocità del flusso aumenta, la grande corona di tentacoli si affloscia, ripiegandosi come un ombrello rovesciato dal vento. Il paragone fra le forze esercitate su un modello di *M. senile* con corona rigida e quelle esercitate su un modello con corona flessibile dimostra che il mutamento passivo di forma riduce notevolmente la resistenza.

Le forze che si esercitano su molte grandi alghe che vivono su scogliere battute dalle onde sono sorprendentemente basse, malgrado le grandi dimensioni di alcune di queste alghe, che grazie alla loro flessibilità possono essere inclinate in una posizione parallela a quella del flusso e, così, venirsi a trovare più vicine al fondo, dove il flusso è più lento. Inoltre, la flessibilità consente alle alghe di mettere in opera un'altra manovra per ridurre gli effetti nocivi delle onde. Dato che le parti mobili si spostano avanti e indietro seguendo il movimento oscillatorio dell'acqua, il flusso di questa rispetto alle alghe è piccolo fino a quando esse non sono completamente ruotate nella direzione del flusso stesso. Quanto più lunga è l'alga, tanto più estesamente viene trascinata dalla corrente prima di disporsi in direzione parallela al flusso. Se un'alga è sufficientemente grande, l'acqua comincerà a scorrere nella direzione opposta prima che l'alga sia stata ruotata completamente. Quindi, per organismi grandi e flessibili un aumento della lunghezza può portare a una diminuzione della resistenza.

Non tutti gli organismi sessili hanno forme che riducono al minimo la resistenza. Per esempio, alcune gorgonie e alcuni alcionari danno origine a colonie a forma di ventaglio che si ergono ad angolo retto rispetto al flusso. Questi animali coloniali si nutrono di materiali sospesi nell'acqua che passa attorno a loro. L'orientamento della colonia in un piano perpendicolare al flusso indica non solo che tutta la colonia è esposta al massimo flusso d'acqua per unità di tempo, ma anche che non ci sono membri della colonia in posizione tale da dover utilizzare un'acqua ormai impoverita delle sostanze nutritive che conteneva. Gordon Leversee, che lavora presso il Duke University Marine Laboratory, ha dimostrato che le colonie piatte delle gorgonie catturano più cibo quando si trovano in posizione perpendicolare al flusso rispetto a quando sono parallele a esso.

In che modo i ventagli di gorgonia assumono l'orientamento ottimale? Stephen A. Wainwright e John Dillon della Duke University hanno trovato che le colonie giovani sono orientate in tutte le direzioni possibili. Essi pensano che la pressione esercitata dal flusso dell'acqua tenda a far ruotare la colonia fino a quando questa non è orientata perpendicolarmente alla direzione del flusso.

Gli organismi marini sessili esposti al flusso delle acque possono anche essere soggetti a forze di sollevamento. Quando la parte superiore dell'organismo si curva verso l'alto è interessata da un flusso d'acqua più veloce di quello che interessa le altre parti. La pressione nella zona di flusso rapido è inferiore a quella esercitata nelle altre regioni, con il risultato che si sviluppa una forza che tende a sollevare l'organismo dal substrato. Mark W. Denny della Stanford University ha potuto misurare forze significative di sollevamento esercitate su vari organismi, come cirripedi e patelle, che si attaccano alle rocce. Analogamente, se la forma di un organismo è tale che l'acqua fluisce più rapidamente da un lato piuttosto che dall'altro, l'organismo sarà spinto verso il lato in cui il flusso è più veloce.

Un'ulteriore forza che si esercita su molti organismi sessili è la reazione all'accelerazione. L'acqua che investe con regolarità una pianta o un animale può aumentare di velocità, rallentare e cambiare direzione. La reazione all'accelerazione spinge l'organismo nella direzione in cui la velocità aumenta. (Si ricordi che, quando l'acqua rallenta, vi è un'accelerazione in senso opposto a quello del movimento.) Maggiore è il volume d'acqua che deve essere accelerato per fluire intorno a un organismo, maggiore è la reazione all'accelerazione. Un grafico che mostra il flusso di un'onda che investe un organismo indica che la resistenza massima e il massimo effetto della reazione all'accelerazione si manifestano in momenti differenti del ciclo dell'onda. La forza alla quale un organismo è sottoposto in ogni istan-

Sono qui illustrate le variazioni di flusso che interessano gli anemoni di mare. Il flusso della corrente principale in un braccio di mare mosso *(a)* raggiunge spesso i cinque metri al secondo quando l'onda si frange sul litorale e durante la risacca. In queste condizioni, l'anemone di mare *Anthopleura xanthogrammica* assume una forma appiattita *(b)* per avvantaggiarsi della minor velocità del flusso in prossimità del fondo. Il valore più elevato di resistenza si ha quando la velocità dell'acqua raggiunge un massimo; la massima reazione all'accelerazione si ha invece quando è massima l'accelerazione. (Il valore dell'accelerazione è alto quando è elevata la pendenza della curva velocità-tempo.) In un sito protetto *(c)*, dove l'impeto delle onde è attutito dalla presenza di scogli, *A. xanthogrammica* assume una forma più alta nella corrente. Alcuni anemoni, come il piccolo *A. elegantissima*, che vivono in gruppi inframmezzati ai più grandi *A. xanthogrammica* in siti protetti *(d)*, sentono ancor meno la forza del flusso.

te è la somma della resistenza e della reazione all'accelerazione in quell'istante.

Se si esaminano le equazioni che esprimono l'entità della resistenza, del sollevamento e della reazione all'accelerazione su organismi sessili, si può notare che le prime due sono proporzionali alla superficie dell'organismo, mentre la terza è proporzionale al volume. Sembra, quindi, che la reazione all'accelerazione assuma un'importanza maggiore, in caso di aumento di dimensioni dell'organismo, rispetto alle altre due forze. Denny e Thomas Daniel della Duke University e io pensiamo che proprio dalla reazione all'accelerazione dipenda il limite massimo delle dimensioni che molti organismi battuti dalle onde possono raggiungere. Certamente molti animali e piante subcotidali, che vivono in correnti più lente e più uniformi di quanto non avvenga per gli organismi intercotidali, raggiungono dimensioni maggiori di questi ultimi.

Può anche darsi che l'azione delle onde determini indirettamente un limite massimo delle dimensioni di certi organismi riducendo il tempo utile a disposizione degli animali per nutrirsi. Suzanne Miller dell'Università di Washington ha scoperto che la forza necessaria per staccare molti gasteropodi marini dalle rocce è inferiore quando gli animali si muovono rispetto a quando stanno fermi. Se la forza delle onde è tale da far diminuire il tempo che un gasteropodo ha a disposizione per muoversi e nutrirsi, la crescita dell'animale sarà limitata. Analogamente, la crescita di un organismo che si nutre di particelle sospese nell'acqua sarà limitata nel caso che un flusso violento non consenta all'organismo stesso di trattenere le particelle di cibo o lo costringa a retrarre le delicate strutture di cui si serve per nutrirsi.

Le forze che si accompagnano al flusso sono in grado, talvolta, di deformare o di rompere un organismo. Il fatto che ciò accada oppure no dipende dall'entità delle sollecitazioni esercitate sui tessuti di un organismo sottoposto al flusso e dalla risposta specifica dei tessuti. Le dimensioni e la forma di un organismo hanno molta importanza nel determinare la risposta all'applicazione di un certo carico. Una pianta o un animale attaccati al substrato possono reagire a un carico in molti modi, ad esempio con uno stiramento longitudinale o trasversale, incurvandosi, o torcendosi.

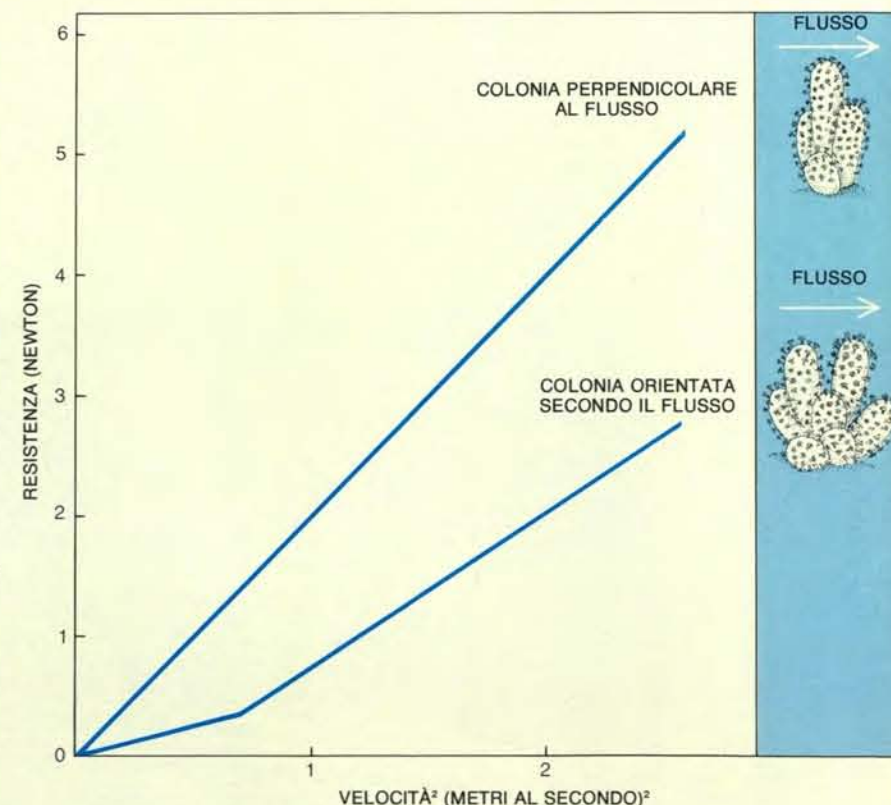
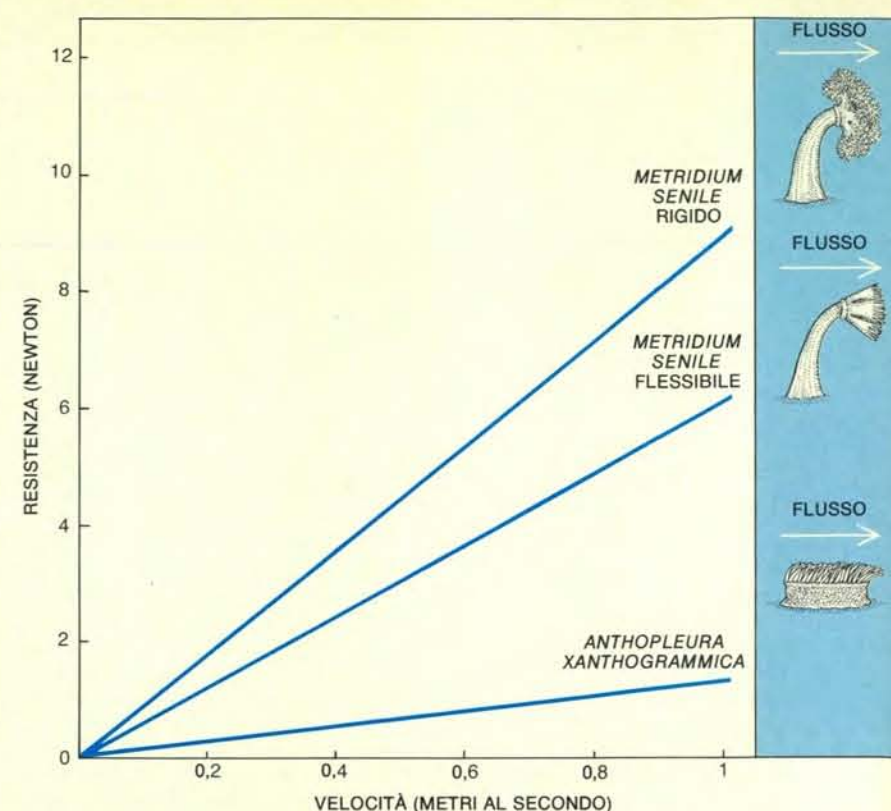
Quando un organismo è sottoposto a forze di taglio o longitudinali, le sollecitazioni esercitate sui suoi tessuti sono inversamente proporzionali alla superficie della sua sezione trasversale, e non dipendono dalla forma di questa né dalla lunghezza del corpo. A parità di carico, le sollecitazioni subite dai tessuti di un organismo di piccole dimensioni sono maggiori di quelle alle quali è sottoposto un organismo di dimensioni maggiori.

Quando un organismo è sottoposto a forze che lo fanno piegare, un lato del suo corpo si stira mentre l'altro si comprime. I tessuti delle superfici opposte sono maggiormente sottoposti rispettivamente a forze di trazione e a forze di compressione. L'entità della sollecitazione dovuta alla trazione e di quella dovuta alla compressione di un punto del corpo dell'organismo sessile risultano inversamente proporzionali alla distanza di quel punto della superficie dell'organismo. Se due organismi dello stesso diametro, ma che differiscono per l'altezza, sono sottoposti allo stesso carico, le sollecitazioni sono maggiori in quello più alto.

L'entità delle sollecitazioni dovute alla trazione e alla compressione sono inversamente proporzionali al cubo del raggio dell'organismo. Pertanto, un lieve aumento del raggio può significare una notevole riduzione delle sollecitazioni. Non sorprende il fatto che molti organismi che tengono una posizione verticale all'interno di un fluido in movimento abbiano una base d'appoggio molto larga e molto rinforzata. Neppure è sorprendente che le sottili costrizioni che si possono trovare nelle strutture di sostegno di organismi sessili svolgano la funzione di giunzioni flessibili. In un organismo sottoposto a forze che ne provocano la torsione, un piccolo incremento del raggio ha come conseguenza un notevole aumento delle sollecitazioni torsionali. Perciò un organismo di piccole dimensioni si torce più prontamente di uno più grande, e la presenza di una costrizione in una fascia del corpo può costituire un punto di rotazione.

Considerando le maggiori sollecitazioni di trazione e di compressione che subisce un organismo il cui raggio si è lievemente ridotto, è evidente che l'asportazione di una parte della struttura di una pianta o di un animale attaccati al substrato in una zona battuta dalle onde può avere effetti fatali. *Nereocystis Luetkeana,* enorme alga bruna del Pacifico nordoccidentale, fornisce un esempio lampante di quanto possa essere disastroso l'effetto di una tale perdita parziale. Una percentuale variabile, a seconda della località, tra il 30 e il 90 per cento delle *Nereocystis* che vengono gettate sulle spiagge hanno uno stipite che si presenta rotto in alcuni punti, in corrispondenza dei quali è stata asportata solamente una piccola quantità di tessuti, generalmente a opera di ricci di mare che li hanno mangiucchiati.

Gli anemoni di mare dimostrano chiaramente quanto le sollecitazioni siano influenzate dalla forma di un organismo. La forza di flusso che si esercita sia su *M. senile* in acque calme sia su *A. xanthogrammica* in bracci di mare mossi è di circa un newton e tende a trascinare gli animali nel senso della corrente. Ho calcolato, però, che per una forza di questo tipo la sollecitazione massima che si esercita nei tessuti di *M. senile,* alto e sottile, è 45 volte maggiore di quella che subisce *A. xanthogrammica,* mediamente basso e largo. In altre parole, *M. senile,* che si trova in un habitat riparato, si trova in condizioni più difficili dal punto di vista delle sollecitazioni meccaniche, rispetto ad *A. xanthogrammica* che vive in un habitat non riparato.

Un organismo può anche ridurre le sol-

Esperimenti condotti su modelli *(figura superiore)* hanno dimostrato che le forze che agiscono su un anemone variano a seconda della configurazione che l'animale assume nei confronti della corrente. La resistenza misurata su un modello di *Anthopleura xanthogrammica* con la maggior parte della superficie parallela al flusso è inferiore a quella che si può misurare su modelli di *Metridium senile*. *A. xanthogrammica* genera una piccola scia, riducendo la resistenza di forma. La resistenza misurata su un modello flessibile di *M. senile* i cui tentacoli, come nell'animale vivente, si raccolgono in una forma che meglio asseconda la corrente a mano a mano che la velocità aumenta, è inferiore a quella misurata su un modello rigido. Le forze che agiscono su un corallo coloniale, come l'*Alcyonium digitatum* della figura inferiore, variano a seconda dell'orientamento della colonia nei confronti del flusso.

lecitazioni sui propri tessuti sostenendo il carico in tensione piuttosto che piegandosi. Un esempio di quanto affermato è dato da due grandi alghe brune che si trovano molto numerose lungo le coste rocciose del Cile. *Lessonia nigrescens* ha una posizione eretta e può piegarsi; *Durvillea antarctica* si trova in posizione inclinata, con il punto di piegamento vicino al rizoide col quale si fissa al fondo, e si orienta secondo il flusso. Per entrambe le specie ho misurato una velocità di flusso fino a sei metri al secondo e forze fino a 20 newton. Il calcolo delle sollecitazioni che si esercitano su stipiti di 20 centimetri di lunghezza e di due centimetri di diametro, soggetti a una forza di 10 newton, indica valori di circa un terzo di newton per metro quadrato se lo stipite viene tirato (come nel caso di *Durvillea*) e di circa 250 newton per metro quadrato se lo stipite viene piegato (come per *Lessonia*).

Il grado di deformazione di un organismo che vive in acque mosse non dipende soltanto dall'entità delle sollecitazioni esercitate sui suoi tessuti, ma anche dalla compattezza di questi, che può venir misurata tirando un frammento di tessuto fino a quando non si rompe (con un apparecchio che ricorda i tavoli di tortura medioevali). È possibile misurare sia la forza che viene esercitata sia la lunghezza raggiunta dal tessuto sottoposto a questa forza. Un grafico dei risultati mostra che occorre una maggior forza per tirare un tessuto compatto sino a fargli raggiungere una certa lunghezza, di quanta ne occorra per un tessuto meno compatto. La pendenza di una curva carico-allungamento rappresenta il coefficiente di elasticità di un materiale e consente di valutare la compattezza del materiale stesso.

Una caratteristica di molti materiali biologici deformabili è che la loro flessibilità dipende dalla velocità con la quale si deformano (cosa che non vale per materiali come l'acciaio o il vetro). Questa caratteristica è alla base di un interessante aspetto del progetto meccanico degli organismi sessili. Un materiale si deforma più rapidamente se viene sottoposto a una sollecitazione notevole di quanto non accada per sollecitazioni ridotte. Dalla forma di un organismo dipende l'entità delle sollecitazioni sui tessuti, e dall'entità delle sollecitazioni dipende il grado di deformazione, che, a sua volta, definisce la compattezza del materiale. Vi è quindi un'importante relazione tra la forma di un organismo e la sua deformabilità.

L'anemone *M. senile* fornisce un esempio della correlazione esistente fra ritmo di vita e deformabilità dei tessuti di un organismo. Un tessuto può essere sottoposto a una prova di tensione, nel corso della quale gli viene applicata una tensione costante e viene misurata la deformazione in funzione del tempo. Gli anemoni di mare posseggono tra la parete esterna e quella interna del corpo un «tessuto» gelatinoso, detto mesoglea. Un grafico d'una prova di tensione eseguita sulla mesoglea di *Metridium* mostra che la mesoglea subisce una piccola deformazione se viene sottoposta a tensione per meno di un minuto. I tempi di deformazione corrispondono a quelli riguardanti i cambiamenti di forma dell'animale dovuti alle cellule contrattili che costituiscono la «muscolatura». La mesoglea, pertanto, rappresenta un supporto abbastanza consistente per l'azione degli elementi contrattili sulle strutture di sostegno dell'organismo.

Consideriamo, ora, l'applicazione di un carico per parecchie ore, come avviene nel caso delle correnti di marea che agiscono sull'anemone. La mesoglea si stira sempre più e l'animale si piega passivamente nel senso della corrente. In un periodo variabile dalle 12 alle 24 ore la mesoglea si può stirare fino a tre volte rispetto alla sua lunghezza originaria, anche per piccoli carichi. Questo lungo periodo corrisponde al tempo necessario a un anemone per gonfiarsi e raggiungere le maggiori dimensioni possibili. (L'animale riesce a far ciò facendo entrare acqua nella cavità gastrovascolare attraverso dei canalicoli detti sifonoglifi, nei quali l'acqua viene fatta scorrere dal battito delle ciglia che li rivestono.) Nei confronti del lavoro veloce delle cellule contrattili la mesoglea si comporta in modo relativamente rigido, ma nei confronti del lavoro lento dei sifonoglifi si dimostra docile.

Un frammento di mesoglea ricavata da *A. xanthogrammica* si comporta in modo assai diverso, sottoposto a una prova di tensione. Infatti questo anemone, anche se viene battuto dalle onde per un'intera giornata in un braccio di mare agitato, non si stira molto. Le mesoglee delle due specie di anemoni rivelano differenze strutturali e molecolari che riflettono l'adattamento alle diverse condizioni meccaniche degli ambienti in cui le due specie vivono.

Le strutture di sostegno di molte piante e animali attaccati al substrato, come il «tronco» degli anemoni di mare e lo stipite delle alghe brune, è formato da più di un materiale. Se il materiale più compatto è più vicino al centro della struttura, questa è più flessibile di quanto sarebbe se tale materiale si trovasse nelle parti più esterne. Stephen Wainwright e io abbiamo studiato un esempio di struttura tanto flessibile da lasciarsi annodare senza deformarsi. Si tratta dello stipite dell'alga bruna *Nereocystis*. Il tessuto vicino al centro dello stipite è più compatto del tessuto alla periferia. Questo fatto non solo pone il tessuto che è in grado di sostenere la maggior quantità di carico in una posizione in cui è difficilmente aggredibile dai ricci di mare, ma fornisce anche all'alga una flessibilità tale da permetterle di disporsi in posizione parallela a una corrente veloce, riducendo la resistenza. Al contrario, i fusti di molte piante terrestri che stanno in posizione eretta (per esempio i girasoli) possiedono elementi cellulari rigidi disposti longitudinalmente nelle parti periferiche.

Le sollecitazioni sui tessuti di un organismo marino sessile sono differenti se l'organismo sopporta i carichi in tensione o se si incurva. Esempi significativi a questo proposito sono forniti da due grandi alghe che formano una densa fascia lungo le coste rocciose del Cile. *Lessonia nigrescens* viene piegata dal flusso dell'acqua in modo tale che un lato del suo stipite viene stirato e l'altro compresso; le massime forze di stiramento e di compressione esercitate su uno stipite sono direttamente proporzionali alla lunghezza di questo e inversamente proporzionali al cubo del suo raggio. *Durvillea antarctica* viene tirata dall'acqua in movimento. La forza di stiramento sullo stipite è indipendente dalla lunghezza di questo ed è direttamente proporzionale al quadrato del suo raggio. Le frecce colorate rappresentano le sollecitazioni.

L'entità delle deformazioni di un organismo sessile nel flusso d'acqua di un determinato ambiente assume una particolare importanza per lo svolgimento delle attività quotidiane da parte dell'organismo stesso. È in grado, questo, di tenere i propri tentacoli in una disposizione adatta alla cattura del cibo? Può mantenere le proprie superfici atte alla fotosintesi al di sopra di quelle di organismi competitori? Corre il rischio di essere danneggiato, o addirittura staccato dal substrato a opera del flusso?

Il fatto che un organismo sessile venga strappato o no dal suo substrato dipende non solo dall'entità delle sollecitazioni che subisce, ma anche dalla resistenza e dalla tenacità dei suoi tessuti e del materiale adesivo che lo tiene fisso al substrato. La robustezza di un materiale è definita dalla forza richiesta per romperlo. Ma la possibilità che qualcosa venga o no strappato da un'onda o da una corrente dipende dalla sua compattezza e dalla sua tenacità ed è definita dal lavoro richiesto per strapparlo. Si può costruire un grafico carico-allungamento per un campione che è stato sottoposto a trazione fino a che non si è rotto; l'area al di sotto della curva rappresenta il lavoro, per volume di materiale, necessario per raggiungere la rottura, cioè una misura della tenacità del materiale.

L'esame d'un gruppo di tali curve mostra che la resistenza alla rottura può venir conseguita in più d'una maniera. Per esempio, il materiale compatto e robusto ottenuto dallo stipite di *Lessonia* non è più tenace di quello ricavato dallo stipite più lasso di *Durvillea*. Queste due alghe illustrano due diverse strategie per resistere alla rottura. Una è quella di essere compatta e robusta, come è *Lessonia;* l'altra è quella di essere gracile e di deformarsi quando sottoposta a un carico, ma d'esser capace di allungarsi notevolmente prima di rompersi, come nel caso di *Durvillea*.

Se una struttura gracile viene sottoposta a carico per un tempo sufficientemente lungo perché l'allungamento raggiunga il punto di rottura, questa si verificherà con un carico minore rispetto a quanto accade per una struttura compatta e robusta. Il modo «debole» di essere resistente ha quindi maggiore efficacia per organismi sottoposti a forze pulsanti di breve durata, come accade su una costa battuta dalle onde. Un organismo gracile, per sopravvivere, deve anche essere dotato di resilienza: deve essere, cioè, in grado di riprendere la propria forma originaria prima dell'arrivo dell'onda successiva. Molte alghe in grado di deformarsi sono anche altamente resilienti. Per esempio, lo stipite di *Nereocystis* è in grado di immagazzinare come energia di deformazione, e di utilizzarla per riacquistare elasticamente la forma primitiva, circa il 75 per cento dell'energia che è stata necessaria per il suo allungamento. Il materiale del suo stipite è resiliente come l'elastina, la proteina elastica che consente alle arterie umane di ritornare al calibro originario dopo esser state deformate dal passaggio dell'onda sfigmica prodotta dal cuore.

Ho progettato delle prove meccaniche per simulare le sollecitazioni che gli organismi marini sessili incontrano in natura. Gli anemoni di mare, specialmente se paragonati ai cirripedi e ai coralli, sono decisamente il gruppo più debole; tuttavia prove di simulazione da me eseguite sulla mesoglea di *A. xanthogrammica* di un braccio di mare mosso indicano che questa mesoglea riprende la sua forma normale onda dopo onda. Sottoposta alle stesse sollecitazioni, la mesoglea di *M. senile*, che vive in acque calme, non ha capacità di recupero altrettanto rapide e, dopo molti cicli, può risultare tanto allungata da rompersi. *M. senile* fa parte di un gruppo di organismi gracili che evidentemente non hanno le caratteristiche di resilienza necessarie per sopravvivere in un braccio di mare mosso.

La robustezza di un tessuto può essere misurata calcolando l'area al di sotto di una curva carico-allungamento di un campione di tessuto sottoposto a trazione fino al punto di rottura. L'area rappresenta il lavoro, per volume di materiale, necessario per rompere il campione. I calcoli mostrano che non c'è una differenza significativa tra le quantità di lavoro (74 chilojoule per metro cubo) necessarie per rompere i tessuti del robusto stipite di *Lessonia*, lo stipite più elastico di *Durvillea* e la lunga fronda della stessa *Durvillea*. Questi esperimenti mettono in evidenza il fatto che gli organismi marini sessili possono mostrare la propria robustezza in maniere diverse.

Molti studiosi di ecologia marina hanno descritto come eventi perturbanti di ordine fisico siano tra i fattori più importanti che influenzano la varietà delle comunità di organismi sessili. I tratti di substrato che restano scoperti quando gli organismi che vivevano su di essi sono stati strappati possono venir colonizzati da piante e animali diversi dai precedenti e che sarebbero stati altrimenti esclusi da quell'ambiente per la competizione con altri organismi più adatti. Chi di noi ha studiato le scogliere coralline della Giamaica prima e dopo che venissero sconvolte dall'uragano Allen nel 1980 ha un'impressione viva di quale ruolo modificatore possa essere svolto da un movimento d'acqua violento su una comunità di organismi. Come ci si può aspettare per quanto ho detto fin qui sulle caratteristiche degli organismi marini sessili, i coralli alti e rigidi delle scogliere vennero mandati in frantumi dall'uragano, mentre le gorgonie merlettate e flessibili e le alghe disposte a tappeto sul fondo sopravvissero, a meno che non fossero state strappate o seppellite dai detriti dei coralli.

Alla ricerca dei numeri primi

Fino a poco tempo fa anche un calcolatore di grandi dimensioni avrebbe impiegato almeno un secolo per stabilire se un numero di cento cifre fosse primo o composto; oggi basta un minuto

di Carl Pomerance

I numeri primi formano la base moltiplicativa del sistema dei numeri. Un numero si dice primo se non è il prodotto di numeri naturali minori di esso. Il numero primo 11, per esempio, non può essere scomposto in fattori minori: solo 1×11 è uguale a 11. Viceversa, un numero si dice composto se può essere espresso come prodotto di due o più fattori. Il numero composto 12 è uguale a $2 \times 2 \times 3$. Ogni intero maggiore di 1 o è primo, o è il prodotto di un unico insieme di numeri primi. Questo fatto, noto già ai greci antichi, ha un ruolo tanto importante, nel sistema dei numeri naturali, da meritare il nome di teorema fondamentale dell'aritmetica.

Com'è possibile determinare se un numero è primo o composto? Il metodo più semplice consiste nel dividere il numero in esame per gli interi della successione 2, 3, 4... Se una qualsiasi delle divisioni non dà resto, il numero è composto e divisore e quoziente sono suoi fattori. Se si sono controllati tutti gli interi fino al numero in questione e nessuna divisione è senza resto, il numero è primo. In effetti, non è necessario arrivare fino al numero in esame; la procedura può essere interrotta non appena il divisore ecceda la sua radice quadrata. La ragione è semplice: i fattori, se esistono, devono essere almeno due e quindi, se un numero ha un fattore maggiore della sua radice quadrata, deve anche averne uno minore.

Limitando le divisioni per tentativi alla radice quadrata, si rende molto più rapido il test di primalità. È possibile mettere in atto anche altre scorciatoie, eliminando per esempio tutti i possibili divisori pari maggiori di 2. Ciononostante, l'algoritmo della serie di divisioni per tentativi è assolutamente inapplicabile al controllo dei più grandi numeri primi conosciuti. Consideriamo il numero $2^{44\,497} - 1$, che Harry L. Nelson e David Slowinski del Lawrence Livermore Laboratory hanno dimostrato essere primo; esso è formato da 13 395 cifre. Se un calcolatore fosse in grado di controllare un milione di divisioni al secondo e si fermasse alla radice quadrata del numero, gli occorrerebbero 10^{6684} anni per espletare il suo compito.

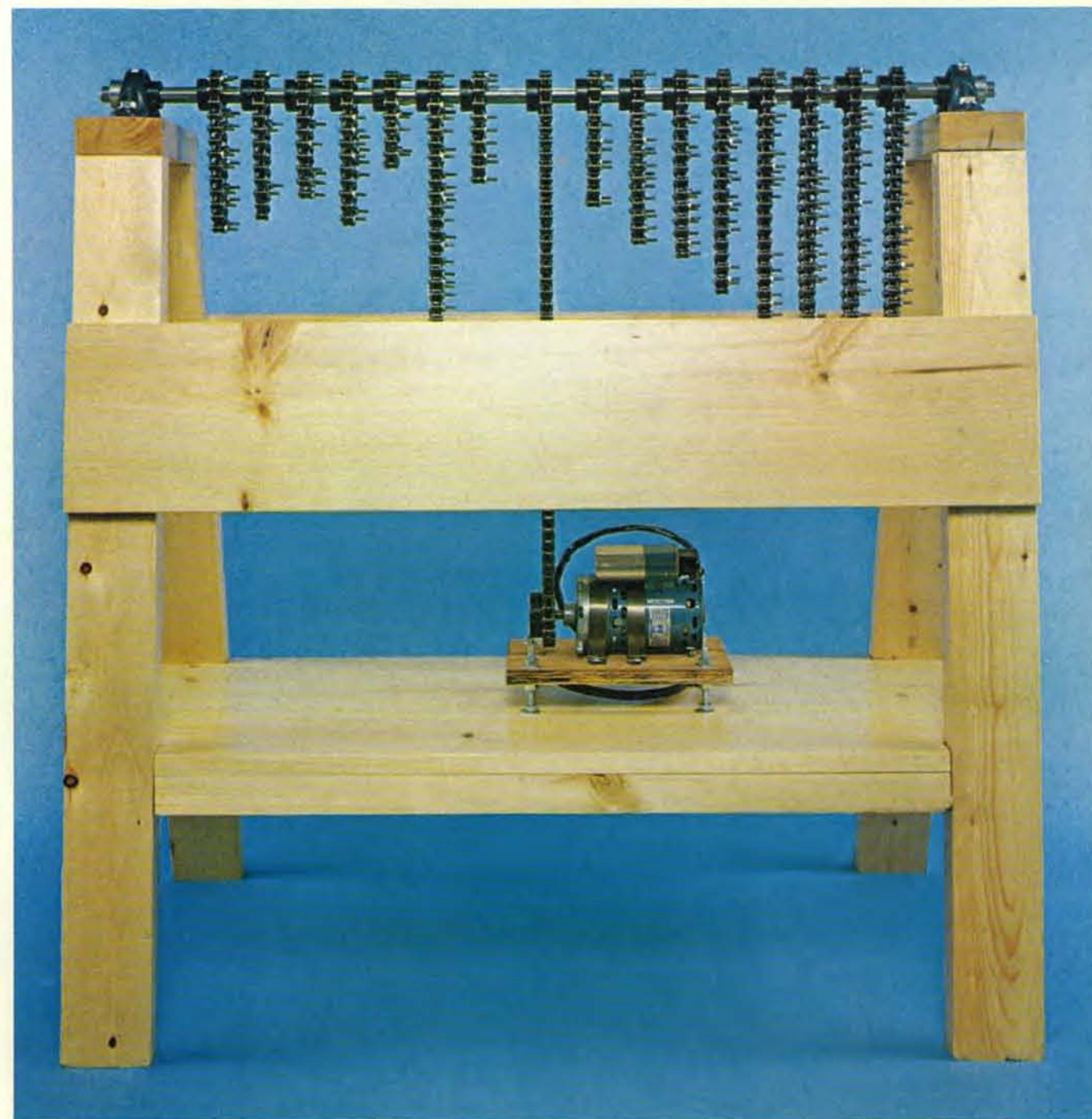
L'inconveniente deriva dal fatto che il metodo delle divisioni fa molto più del necessario: non solo ci dice se un numero è primo o composto, ma ci fornisce anche i fattori di ogni numero composto. Sebbene esistano metodi di fattorizzazione indipendenti da quello delle divisioni per tentativi, nessuno di essi è in grado di fattorizzare in un tempo ragionevole un numero arbitrario di «sole» 100 cifre, anche ricorrendo a un grande calcolatore. Del resto, è possibile stabilire se un numero è primo o no senza essere costretti a trovarne i fattori qualora sia composto. Se il numero non possiede fattori piccoli, questi metodi risultano quasi sempre più efficienti di quelli che usano la fattorizzazione. Negli ultimi anni è stato sviluppato un metodo che consente al calcolatore di determinare la primalità di un qualsiasi numero di 100 cifre in 40 secondi di tempo macchina.

Il problema della primalità e quello a esso superficialmente collegato della fattorizzazione appartengono alla teoria dei numeri, quella branca della matematica che tratta delle proprietà dei numeri interi; essa è ricca di problemi straordinariamente semplici nella loro formulazione, ma notoriamente difficile da risolvere. I problemi di teoria dei numeri che hanno a che fare con la primalità hanno affascinato i matematici fin dai tempi di Euclide.

Così pare che esista un'infinità di numeri primi gemelli, ossia coppie di primi che, al pari di 17 e 19, differiscono di 2. Ma si tratta di una congettura non dimostrata. È quasi certamente vero che esiste sempre almeno un numero primo fra i quadrati di due numeri successivi, ma anche questa è un'asserzione non dimostrata. Christian Goldbach nel 1742 ipotizzò che ogni numero pari maggiore di 2 sia la somma di due primi; 32, per esempio, è la somma di 13 e 19. La congettura di Goldbach ha resistito a tutti i tentativi di dimostrazione, anche se nel 1937 il matematico russo I. M. Vinogradov ha dimostrato che tutti i numeri dispari «abbastanza grandi» possono essere espressi come somma di tre primi. Vinogradov non è stato tuttavia in grado di esplicitare chiaramente cosa si debba intendere per «abbastanza grandi».

Un corollario del teorema di Vinogradov asserisce che tutti i numeri pari abbastanza grandi possono essere espressi come somma di quattro primi. In base al teorema, poi, se a è un numero pari abbastanza grande e b è un numero dispari tale che $b = a - 3$, è sempre possibile rappresentare b come somma di tre primi. Il numero pari a in questione è quindi la somma di quei tre primi e di 3 (numero primo). Nel 1966 il matematico cinese Chen Jing-run mostrò che tutti i numeri pari abbastanza grandi possono essere espressi come somma di un numero primo e di un numero che o è primo, o è la somma di due primi. Queste approssimazioni alla congettura di Goldbach costituiscono risultati profondi, nel senso che la loro dimostrazione richiede un'analisi matematica avanzata ed è piuttosto difficile.

Molti enunciati sui numeri primi, però, possono essere dimostrati con metodi elementari, e molte di queste dimostrazioni sono deliziosamente ingegnose. Così, era noto già a Euclide che il numero dei primi è infinito. L'argomentazione è di tipo indiretto; si assume che il numero dei primi sia finito, dal che scende l'esistenza di un massimo numero primo, e successivamente si ricava una contraddizione. Supponiamo allora che p sia il massimo numero primo e consideriamo il numero N definito come il prodotto di tutti i primi fra 2 e p. Il numero $N + 1$ sarà primo o composto. Dato che $N + 1$ è maggiore di p, in base all'ipotesi di partenza dovrà essere composto (altrimenti sarebbe un primo maggiore di p); ma allora per il teorema fondamentale dell'aritmetica dovrà avere dei fattori primi. Peraltro $N + 1$ era stato costruito in modo che la sua divisione per un numero primo qualunque, compreso fra 2 e p, desse sempre resto 1. I suoi fattori primi (se esistono) devono quindi essere maggiori di p. L'ipotesi che esista un massimo numero primo conduce allora a una contraddizione, sicché l'insieme dei primi non ha un limite superiore.

In modo analogo è facile dimostrare che esistono dei primi consecutivi la cui differenza può essere grande quanto si vuole. Consideriamo la successione di numeri $n! + 2, n! + 3, n! + 4, ..., n! + n$, dove $n!$ (da leggersi n fattoriale) è il prodotto di tutti gli interi da 1 a n. Si noti che $n! + 2$ è divisibile esattamente per 2, $n! + 3$ per 3, $n! + n$ per n. Tutti gli $n - 1$ numeri della successione sono composti e la suc-

La prima macchina approntata per lo studio del sistema numerico fu costruita nel 1926 da D. H. Lehmer, dell'Università della California a Berkeley. Costruita con un cavalletto di legno, catene di bicicletta e altro materiale facilmente reperibile, questa macchina era una sorta di calcolatore specializzato, che poteva essere programmato per la ricerca rapida di numeri di forma particolare, necessari per la risoluzione di determinati problemi di teoria dei numeri. Il test di primalità, ovvero la classificazione di un numero come primo o composto è uno fra i più importanti di questi problemi. (Un numero primo è un numero divisibile esattamente solo per se stesso e per 1; se un numero possiede altri divisori, è composto.) Le condizioni che devono essere soddisfatte da una soluzione numerica del problema devono essere programmate sulla macchina di Lehmer inserendo dei bulloni in certi anelli delle catene di bicicletta. Quando queste vengono fatte girare da un motore, la macchina rimane in movimento finché tutti i bulloni non sono allineati in alto; quindi si spegne automaticamente. Il numero corrispondente alla configurazione delle catene al momento dell'arresto soddisferà le condizioni imposte. Lehmer costruì diverse versioni della macchina, ma la versione del 1926 è andata distrutta. La macchina mostrata nella fotografia è una ricostruzione dovuta a Roberto Canepa della Carnegie-Mellon University, conservata al Computer Museum di Marlboro, nel Massachusetts.

cessione stessa può essere prolungata a piacere scegliendo n abbastanza grande.

Molti matematici hanno considerato la teoria dei numeri come la «regina della matematica», sia per l'intricata bellezza delle sue dimostrazioni, sia perché il suo studio ha sempre avuto il fascino di una forma di pura contemplazione, senza il peso di possibili conseguenze pratiche. Dal 1977, tuttavia, lo sviluppo della teoria dei numeri è stato stimolato anche dalla scoperta che essa avrebbe potuto avere importanti applicazioni nella crittografia, lo studio delle condizioni di riservatezza nelle comunicazioni. In quell'anno, Ronald L. Rivest del Massachusetts Institute of Technology, Adi Shamir del Weizmann Institute of Science e Leonard M. Adleman dell'Università della Southern California notarono che era possibile fondare un sistema crittografico a chiave pubblica sulla difficoltà di fattorizzazione di un numero composto molto grande, che sia per esempio il prodotto di due numeri primi di 100 cifre.

In un sistema crittografico a chiave pubblica lo strumento di codifica del messaggio può essere reso di ragione pubblica senza pregiudicare la sicurezza del codice. Il codice Rivest-Shamir-Adleman si basa sulla relativa facilità della determinazione della primalità di due grandi numeri e del valore del loro prodotto da un lato, e dalla elevata difficoltà pratica di fattorizzare tale valore se non si sa come è stato costruito. Rendendo pubblico il prodotto di 200 cifre di due primi di 100 cifre, chiunque potrebbe codificare un messaggio utilizzando tale numero, ma solo la conoscenza dei due fattori primi renderebbe possibile la decodifica. Esistono sistemi crittografici a chiave pubblica che non dipendono dalla fattorizzazione, ma il sistema Rivest-Shamir-Adleman basa la sua sicurezza sulla intrattabilità del problema della fattorizzazione e la sua operatività sul fatto che i due numeri di 100 cifre siano effettivamente primi. Un test di primalità che non dipenda dalla fattorizzazione sarebbe quindi di grande interesse per i sistemi crittografici (si veda l'articolo *Crittografia a chiave pubblica* di Martin E. Hellman, in «Le Scienze», n. 136, dicembre 1979).

Il crivello per la ricerca dei numeri primi, attribuito all'antico studioso greco Eratostene, è stata una delle prime tecniche escogitate per distinguere i numeri primi, compresi entro un certo limite, da quelli composti. Qui il crivello è un piano inclinato in cui sono stati praticati dei fori. I numeri da controllare rispetto alla primalità sono rappresentati da biglie che rotolano lungo scanalature sul piano. Dapprima si fanno dei fori nella seconda riga, una scanalatura sì e una no, all'infuori di quella contrassegnata con 2. Quindi si cerca la prima scanalatura senza fori, in pratica la terza. Si eseguono dei fori nella terza riga, ogni tre scanalature, all'infuori della scanalatura contrassegnata con 3. Il procedimento continua con l'approntamento di fori nella quinta riga ogni cinque scanalature, ogni sette scanalature nella settima riga, e via dicendo, fermandosi alla riga contrassegnata da un numero minore o uguale alla radice quadrata del massimo dei numeri da controllare. Le biglie che non cadono nei fori corrispondono ai numeri primi. Per esempio, tutti i 25 primi minori di 100 possono essere determinati considerando tutte le biglie da 2 a 100 che superano la settima riga di fori. (Non esistono primi maggiori di sette minori o uguali alla radice quadrata di 100.)

Tutti i metodi di controllo della primalità che non dipendono dalla fattorizzazione traggono la loro importanza da un teorema formulato per la prima volta in una lettera inviata da Pierre de Fermat all'amico Bernard Frénicle de Bessy, datata 18 ottobre 1640. Il teorema, noto come piccolo teorema di Fermat, stabilisce che se n è un numero primo e b un intero, $b^n - b$ è un multiplo di n. Per esempio, se n è uguale a 7 e b è uguale a 2, il teorema asserisce correttamente che $2^7 - 2$, ovvero 126, è un multiplo di 7.

Rispetto alla primalità, l'importanza del teorema sta nella sua equivalenza logica con l'asserzione che se $b^n - b$ non è un multiplo di n, allora n è un numero composto. Quando n è uguale a 4 e b è uguale a 3, $3^4 - 3$ è uguale a 78, che dà resto diverso da zero (per la precisione, 2) quando viene diviso per 4. Il piccolo teorema consente di concludere in modo indiretto che 4 non è primo.

L'aritmetica modulare costituisce un sistema di calcolo con importanti applicazioni ai test di primalità. Nell'aritmetica modulare l'unica cosa che conta in un numero n è il resto che si ottiene dividendo n per un qualche modulo m. La grandezza assoluta di un numero viene completamente ignorata. Il più familiare dei sistemi di aritmetica modulare è quello utilizzato per il calcolo del tempo, in cui le ore vengono indicate specificandone il valore modulo 12. Il simbolo ≡ va letto «è congruo a»; i numeri ai suoi lati danno lo stesso resto quando vengono divisi per il modulo. Per esempio, 4 ≡ 1 (mod 3) significa che 4 e 1 danno lo stesso resto (1) quando vengono divisi per 3.

Per quanto il piccolo teorema costituisca un risultato fondamentale e potente, possiede diverse dimostrazioni elementari, una delle quali verrà riportata più sotto. Il teorema consente di stabilire proprietà di numeri così grandi da non poter essere scritti in forma decimale. Dal fatto che $2^{44\,497} - 1$ è primo, per esempio, il teorema fa scendere che il numero ottenuto elevando 3 alla $(2^{44\,497} - 1)$-esima potenza e togliendo dal risultato 3, è ancora divisibile per $2^{44\,497} - 1$. Il risultato dell'elevazione a potenza è tanto grande da non poter essere scritto sotto forma decimale. Inoltre, il processo di divisione che dovrebbe fornire esplicitamente il quoziente non potrebbe essere eseguito da alcun calcolatore fisicamente concepibile.

Per arrivare ad avamposti così lontani del sistema dei numeri, si può usare la ruota aritmetica ideata da Carl Friedrich Gauss. A lui si deve la formulazione dell'aritmetica modulare, in cui la grandezza assoluta di un numero è irrilevante e ciò che conta è la dimensione dell'ultimo giro compiuto dalla ruota aritmetica per raggiungerlo. Il numero n viene espresso come resto dopo essere stato diviso per un qualche altro numero m, detto modulo. Il resto viene scritto nella forma n modulo m, oppure n (mod m). Il numero m indica la grandezza della ruota, n la grandezza assoluta del numero e n (mod m) la grandezza dell'ultimo giro parziale della ruota necessario per raggiungere n.

Nell'aritmetica modulare esistono analoghi stretti di molte leggi dell'aritmetica ordinaria. In particolare è possibile sommare e moltiplicare, a condizione di esprimere i risultati come congruenze; tutti i numeri che danno uno stesso resto rispetto a un dato modulo sono detti congrui rispetto a quel modulo.

Nell'aritmetica ordinaria, 6 più 7 è uguale a 13, un risultato riproducibile in aritmetica con un modulo, per esempio, 5. Si ha infatti che 6 (mod 5) + 7 (mod 5) è congruo a 1 + 2 ovvero a 3, e che 13 (mod 5) è anch'esso congruo a 3. Analogamente, 4 × 5 è uguale a 20, mentre nell'aritmetica modulo 3 il prodotto ha la forma 4 (mod 3) × 5 (mod 3), e il risultato è congruo a 1 × 2. Tale prodotto è quindi 2, ma anche 20 (mod 3) è congruo a 2.

Nella notazione di Gauss il piccolo teorema di Fermat stabilisce che, se n è primo, $b^n - b$ è congruo a 0 (mod n) o, in altri termini, $b^n - b$ è un multiplo di n. Il vantaggio di questa notazione sta nel fatto che le regole dell'aritmetica modulare rendono possibile calcolare il valore di $b^n - b$ modulo n senza dover dividere $b^n - b$ per n. Per numeri come $2^7 - 2$ il vantaggio offerto dal sistema di Gauss non sembra molto significativo, in quanto la divisione diretta è semplice. Tuttavia per trovare il resto, quando si divide un numero come $3^{1037} - 3$ per 1037, l'aritmetica modulare diventa pressoché indispensabile.

L'essenza del problema sta nel trovare 3^{1037} (mod 1037); nell'aritmetica modulare non è necessario calcolare il valore dell'enorme numero 3^{1037}. Tutto ciò che serve è la ripetuta applicazione del fatto che il resto del quadrato di un numero è congruo al quadrato del resto del numero.

Così, una volta calcolato 3^8 (mod 1037), si può ottenere 3^{16} (mod 1037) elevando al quadrato il resto di 3^8 e trovando il resto di questo numero modulo 1037. In questo modo si possono trovare i resti modulo 1037 di 3, 3^2, 3^4... fino a 3^{1024}. Il numero 3^{1037} è uguale a $3^{(1024+8+4+1)}$, che a sua volta è uguale a $3^{1024} \times 3^8 \times 3^4 \times 3$ per la legge del prodotto di potenze della stessa base; quindi 3^{1037} (mod 1037) è congruo a 3^{1024} (mod 1037) × 3^8 (mod 1037) × 3^4 (mod 1037) × 3 (mod 1037). Quando si sono eseguiti tutti i calcoli, si trova che 3^{1037} è congruo a 845 (mod 1037), per cui $3^{1037} - 3$ è congruo a 842 (mod. 1037) *(si veda l'illustrazione a pagina 91)*. Sulla base del piccolo teorema di Fermat, 1037 deve quindi essere composto, dato che il resto della divisione di $3^{1037} - 3$ per 1037 non è uguale a zero. Questo procedimento non dà praticamente alcun suggerimento per l'identificazione dei fattori di 1037.

Può essere divertente seguire il procedimento con un calcolatore programmabile. Per evitare errori di arrotondamento, i valori di n dovrebbero essere limitati a numeri le cui cifre siano al più la metà di quelle visualizzabili sul calcolatore. Con un calcolatore di grandi dimensioni, i calcoli possono essere eseguiti rapidamente anche quando i numeri in ingresso hanno migliaia di cifre, quindi con il test di Fermat si possono individuare numeri composti enormemente grandi.

La dimostrazione del piccolo teorema di Fermat deriva da una semplice conseguenza del teorema fondamentale dell'aritmetica. Se un numero primo divide esattamente il prodotto di più numeri, divide esattamente almeno uno dei fattori. 4 × 9, ossia 36, per esempio, è divisibile esattamente per il primo 3 e ovviamente anche uno dei suoi fattori, 9, è divisibile per 3. L'enunciato non vale per numeri composti: 4 × 9 è divisibile per 6, ma né 4, né 9 sono divisibili per 6.

Per dimostrare il risultato di Fermat secondo cui $b^n - b$ è un multiplo di n quando è primo, notiamo che $b^n - b$ è uguale a $b\,(b^{n-1} - 1)$. Quindi, se b stesso è un multiplo di n lo è anche $b^n - b$. Il

teorema resta da dimostrare solo per il caso in cui b non sia un multiplo di n: la dimostrazione parte da questa ipotesi.

L'idea di base è che se i numeri b, $2b$, $3b$... fino a $(n-1)b$ vengono moltiplicati fra loro, il loro prodotto può essere espresso nella forma $b^{n-1}(n-1)!$. D'altra parte, dal teorema fondamentale dell'aritmetica segue che i resti modulo n di b, $2b$, $3b$... $(n-1)b$ sono i numeri 1, 2, 3... $(n-1)$ eventualmente in ordine differente. Un po' di algebra elementare ci consente di concludere che $(b^{n-1}-1)(n-1)!$ è un multiplo di n. Dato che il numero primo n non è un divisore di nessuno dei numeri fra 1 e $n-1$, mentre è un divisore di $(b^{n-1}-1)(n-1)!$, un'altra applicazione del teorema fondamentale dell'aritmetica comporta che n è un divisore di $b^{n-1}-1$. Poiché $b^{n-1}-1$ è un fattore di (b^n-b), il teorema è dimostrato.

Potrebbe sembrare che il piccolo teorema di Fermat risolva completamente il problema della primalità, poiché sembra fornire un metodo semplice per distinguere un numero primo da uno composto; sfortunatamente si tratta di una conclusione affetta da un vizio logico. Se per qualche numero b, b^n-b dà resto differente da zero quando viene diviso per n, allora n è certamente composto. Supponiamo tuttavia che b^n-b sia un multiplo di n. Ne segue che n deve essere primo?

Svariati esempi suggeriscono una risposta affermativa: 2^2-2 è un multiplo di 2, 2^3-2 è multiplo di 3, 2^5-2 è multiplo di 5 e 2, 3, 5 sono tutti primi. Circa 2500 anni fa i matematici cinesi scoprirono tale campione e ipotizzarono che se 2^n-2 è multiplo di n, n deve essere primo. Gottfried Wilhelm Leibniz, che svolse uno studio sulle figure binarie dello *I Ching*, prestò fede a questo risultato. Nel 1919, tuttavia, il matematico francese Pierre Frèdéric Sarrus notò che $2^{341}-2$ è un multiplo di 341 anche se 341 è un numero composto, il prodotto di 11 e 31. Dal lavoro di Sarrus sono stati scoperti molti altri controesempi riguardanti diversi valori della base b: $3^{91}-3$ è un multiplo del numero composto 91, $4^{15}-4$ è un multiplo del numero composto 15. Tutti questi enunciati possono essere controllati con un piccolo calcolatore sfruttando l'aritmetica modulare già descritta.

Un numero che, secondo il test di Fermat per un valore assegnato di b, non risulta composto, prende il nome di pseudoprimo in base b. Il numero 341 è pseudoprimo in base 2, mentre 91 è pseudoprimo in base 3 e 15 in base 4. Di fatto esiste un'infinità di pseudoprimi in ogni base b. Alcuni numeri composti, come 561 (il prodotto di 3, 11, 17) e 1729 (il prodotto di 7, 13, 19) sono pseudoprimi in ogni base b. Numeri siffatti sono detti numeri di Carmichael dal nome del matematico americano R. D. Carmichael, che scoprì le loro proprietà nel 1909.

L'esistenza di numeri di Carmichael pone fine alla speranza che il test di Fermat, almeno nella sua formulazione originaria, possa separare tutti i primi dai composti. Peraltro, i numeri di Carmichael sono molto rari, e rari sono anche gli pseudoprimi in una singola base, se confrontati con i numeri primi. Jan Bohman dell'Università di Lund ha dimostrato che esistono esattamente 882 206 716 primi minori di 20 miliardi. John L. Selfridge della rivista «Mathematical Reviews» e Samuel S. Wagstaff, Jr. dell'Università della Georgia hanno calcolato che esistono solo 19 865 pseudoprimi in base 2 minori di 20 miliardi. Se si eseguisse il test di Fermat in base 2 per tutti i numeri inferiori a 20 miliardi, la percentuale di errore sarebbe di circa uno su un milione.

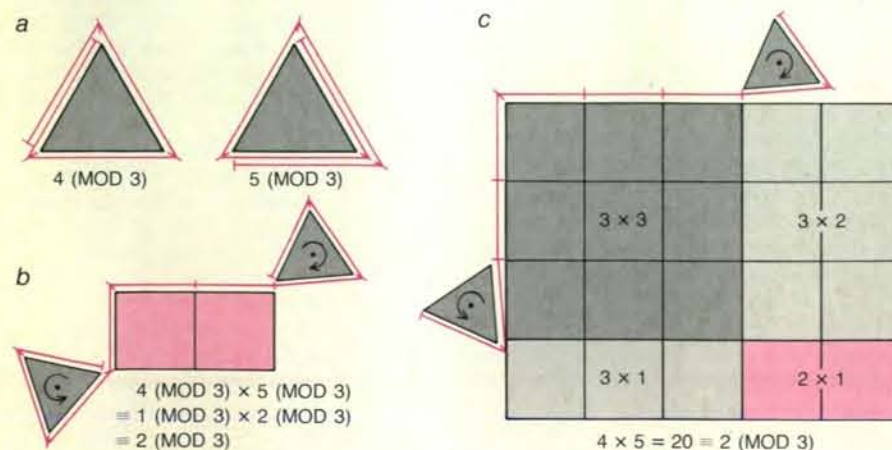
La scarsità di pseudoprimi in base 2, fra tutti i numeri inferiori a 20 miliardi, fa pensare che, se un numero supera il test di Fermat in base 2, è molto probabile che sia primo. Inoltre, se il numero è un numero composto che supera il test di Fermat in base 2, potrebbe non passarlo in base 3. Sarebbe piacevole poter asserire che, applicando il test di Fermat alla base 3, si riduce significativamente la probabilità che un numero composto continui a comportarsi come se fosse un numero primo. Poiché tuttavia i test possono non essere indipendenti, è possibile che quello in base 3 non elimini ancora molti composti che non siano già stati eliminati dal test in base 2.

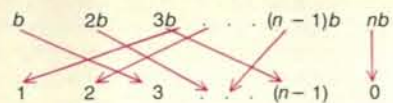
Recentemente D. H. Lehmer dell'Università della California a Berkeley e, indipendentemente, Robert M. Solovay del California Institute of Technology e Volker Strassen del Politecnico federale svizzero hanno sviluppato una variante del test di Fermat che soddisfa al requisito dell'indipendenza in differenti basi. Il test ha la caratteristica che se il numero n in esame è composto, risulta tale per almeno la metà dei valori della base b compresi fra 1 e n. Sicché scegliendo a caso 100 basi differenti e applicando il test di Lehmer-Solovay-Strassen, la probabilità che qualche numero composto n superi tutti e 100 i test è minore o uguale a una su 2^{100} ovvero a una su 10^{30}.

Solovay e Strassen dicono che il loro è un «test di primalità Monte Carlo» (numerosi metodi probabilistici matematici e fisici prendono il nome da quella città, nota per i suoi giochi d'azzardo). Dal punto di vista pratico la denominazione sembra precisa; parrebbe infatti che nella applicazione crittografica il funzionamento di un codice non venga inficiato dalla piccolissima probabilità che i numeri su cui si basa non siano in realtà primi. È stato anche osservato, su basi che ritengo filosoficamente fragili, che dal momento che ogni ordinaria dimostrazione è soggetta a correzioni ed errori umani, si potrebbe accettare una forte conferma probabilistica della primalità come una sua dimostrazione matematica.

In effetti abbiamo buone ragioni per credere che la probabilità di errore di una dimostrazione matematica sia molto superiore a 1 su 10^{30}. La storia della matematica offre numerosi esempi di «dimostrazioni» che si sono poi rivelate erronee o fuorvianti. Esiste però una differenza qualitativa, di grande rilievo per i matematici, fra una conferma probabilistica e una dimostrazione matematica. Una dimostrazione è un argomento deduttivo, in cui ogni passo segue dai precedenti. La forza della dimostrazione non deriva solo dal fatto che è possibile controllare la validità delle conclusioni, ma anche dal fatto che la validità segue in forza dell'argomento. Ritengo che il test di primalità Monte Carlo induca piuttosto a pensare che fra i concetti di dimostrazione e di certezza corre una notevole differenza.

Nel 1876 il matematico francese Edouard A. Lucas espose un solido test di primalità per un qualsiasi numero n. Supponiamo che esista un numero b per il quale b^{n-1} sia congruo a 1 modulo n, mentre $b^{(n-1)/p}$ non sia congruo a 1 modulo n per ogni fattore primo p di $n-1$. Lucas dimostrò che in tal caso n deve essere primo.

Nell'aritmetica modulare l'addizione viene eseguita come nell'aritmetica ordinaria. Si determinano i resti di ciascun numero rispetto al modulo assegnato e successivamente li si somma. Se il valore ottenuto è maggiore del modulo, si trova il resto della somma. Nell'esempio viene calcolata la lunghezza totale di due corde, rispettivamente pari a 6 e 7, modulo 5. Nell'aritmetica modulo 5 si ignora il numero degli avvolgimenti completi attorno a un pentagono, e si considera unicamente quella parte della corda che resta dopo l'ultimo giro completo. Quindi, modulo 5, 6 è congruo a 1 e modulo 7 a 2 *(a)*. Quando si considerino unite le due corde, la lunghezza totale dei due resti è 3 *(b)*. Avvolgendo la corda attorno al pentagono, la sua lunghezza eccedente i giri completi è 3, di modo che la somma ordinaria di 6 e 7 è congrua a 3 modulo 5 *(c)*. Come regola generale, la somma dei resti di due numeri è congrua al resto dell'ordinaria somma degli stessi.

Nell'aritmetica modulare anche la moltiplicazione viene data in modo analogo a quella ordinaria. Nell'esempio, 4 e 5 sono moltiplicati modulo 3 e vengono rappresentati come corde avvolte intorno a un triangolo. Il resto di ogni numero è la lunghezza della corda che resta dopo l'ultimo giro completo attorno al triangolo *(a)*. Il resto di 4 modulo 3 è 1 e il resto di 5 modulo 3 è 2. Il prodotto dei due resti è l'area del rettangolo che ha lati uguali alla lunghezza di ciascun resto. In altri termini, 1 per 2 è uguale a 2 *(b)*. Il prodotto di 4 e 5, d'altra parte, è l'area del rettangolo di lati di lunghezza 4 e 5. Il resto del prodotto *(c)* viene ottenuto ignorando l'area di ciascun rettangolo piccolo i cui lati abbiano lunghezza pari a quella di una corda che compia un numero intero di avvolgimenti completi attorno al triangolo *(aree in grigio)*. L'area del rettangolo restante *(in colore)* è il resto del prodotto di 4 per 5 modulo 3. La regola generale asserisce pertanto che il prodotto dei resti di due numeri è congruo al resto del loro prodotto.

Con l'aritmetica modulare si evitano calcoli che nell'aritmetica ordinaria richiederebbero molto tempo e sarebbero esposti a errori. Il diagramma di flusso mostra come si possa trovare il resto quando si divide il numero 3^{1037} per 1037, senza dover calcolare il valore effettivo di 3^{1037}. Il metodo si fonda sul fatto che il resto del quadrato di un numero per un dato modulo è uguale al quadrato del resto di quel numero per lo stesso modulo. Elevando ripetutamente al quadrato il resto di una potenza di 3 e prendendo il resto del risultato, si può trovare una successione di resti modulo 1037, di 3, 3^2, 3^4, 3^8, 3^{16}, fino a 3^{1024}. Dato che 3^{1037} è uguale a $3^{1024} \times 3^8 \times 3^4 \times 3$, il resto di 3^{1037} diviso per 1037 è uguale al resto del prodotto dei resti 3^{1024} (mod 1037) $\times$ 3^8 (mod 1037) $\times$ 3^4 (mod 1037) $\times$ 3 (mod 1037). Il procedimento può essere svolto da un calcolatore programmabile.

La dimostrazione del «piccolo teorema» di Fermat discende dall'applicazione del teorema fondamentale dell'aritmetica e dalle regole di moltiplicazione nell'aritmetica modulare. Il piccolo teorema di Fermat asserisce che se n è un numero primo e b un intero, $b^n - b$ è un multiplo di n. Se b è un multiplo di n, la conclusione è pressoché immediata: dato che $b^n - b$ è uguale a $b(b^{n-1} - 1)$, b è un fattore di $b^n - b$ e $b^n - b$ è un multiplo di n. Se b non è multiplo di n, basta dimostrare che $b^{n-1} - 1$ è multiplo di n. Un corollario del teorema fondamentale dell'aritmetica ci dice che se un primo divide esattamente il prodotto di più numeri, divide anche uno dei numeri. Quindi, poiché n non divide esattamente nessuno dei numeri fra 1 e $n - 1$, non dividerà esattamente neppure $(n - 1)!$. Ne segue che se si riesce a dimostrare che n divide esattamente $(b^{n-1} - 1)(n - 1)!$, allora anche $b^{n-1} - 1$ e $b^n - b$ sono divisibili esattamente per n; ne consegue che $(b^{n-1} - 1)$ e $(b^n - b)$ sono anch'essi divisibili per n. Consideriamo i numeri b, $2b$, $3b$, fino a nb. Nessuna coppia di numeri ib e jb può dare lo stesso resto quando siano divisi per n. Altrimenti $ib - jb$, che è uguale a $(i - j)b$, sarebbe un multiplo di n, dato che la sottrazione cancellerebbe i resti. Poiché b non è multiplo di n, il teorema fondamentale dell'aritmetica implica che $i - j$ è un multiplo di n. Il numero n, tuttavia, non può essere diviso senza resto da un qualsiasi numero $i - j$, dove i e j siano stati scelti nella successione di numeri da 1 a n. La supposizione che ib e jb diano lo stesso resto, una volta divisi per n, conduce a una contraddizione. Poiché il resto della divisione di nb per n è 0 e la divisione per n dà resti solo fra 1 e $n - 1$, i numeri b, $2b$,..., $(n - 1)b$, se divisi per n danno resti compresi fra 1 e $n - 1$, in un ordine opportuno. Il diagramma mostra che dividendo per n i numeri b, $2b$,... $(n - 1)b$ e prendendo i resti, si può stabilire una corrispondenza fra b, $2b$,... $(n - 1)b$ e 1, 2,... $(n - 1)$. Dato che i resti dei numeri di ciascun insieme sono gli stessi, a meno dell'ordine, il resto del prodotto $b \times 2b \times ... \times (n \times 1)b$ è uguale al resto del prodotto $1 \times 2 \times ... \times (n - 1)$. Sottraendo un prodotto dall'altro si cancellano i resti e la differenza dei due prodotti è un multiplo di n. Con pochi passaggi algebrici è possibile dimostrare che la differenza dei due prodotti è uguale a $(b^{n-1} - 1)(n - 1)!$, per cui l'espressione è divisibile esattamente per n. È stato in questo modo dimostrato il piccolo teorema di Fermat per tutti i numeri interi b.

341 PSEUDOPRIMO IN BASE 2

$2^{341} = 2^{256} \times 2^{64} \times 2^{16} \times 2^4 \times 2$
$\equiv 64 \times 16 \times 64 \times 16 \times 2$ (MOD 341)
$\equiv 2$ (MOD 341)
QUINDI $2^{341} - 2 \equiv 0$ (MOD 341)
PERTANTO 341 SUPERA IL TEST DI FERMAT IN BASE 2
MA $341 = 11 \times 31$

561 PSEUDOPRIMO IN OGNI BASE

$2^{561} = 2^{512} \times 2^{32} \times 2^{16} \times 2$
$\equiv 103 \times 103 \times 460 \times 2$ (MOD 561)
$\equiv 2$ (MOD 561)
QUINDI $2^{561} - 2 \equiv 0$ (MOD 561)
PERTANTO 561 SUPERA IL TEST DI FERMAT IN BASE 2
MA $561 = 3 \times 11 \times 17$
$3^{561} = 3^{512} \times 3^{32} \times 3^{16} \times 3$
$\equiv 273 \times 273 \times 69 \times 3$ (MOD 561)
$\equiv 3$ (MOD 561)
QUINDI $3^{561} \times 3 \equiv 0$ (MOD 561)
PERTANTO 561 SUPERA IL TEOREMA DI FERMAT IN BASE 3

Gli pseudoprimi sono numeri che superano il test di primalità derivato dal piccolo teorema di Fermat per qualche base b, ma che tuttavia sono numeri composti. Quindi, uno pseudoprimo in base b è un numero composto n che divide senza resto $b^n - b$. Il matematico francese Pierre Frédéric Sarrus osservò per primo nel 1919 che 341, il prodotto di 11 per 31, è uno pseudoprimo in base 2. Nel nostro caso viene mostrato che $2^{341} - 2$ è divisibile per 341, sfruttando l'aritmetica modulare in modo analogo a quello dell'illustrazione della pagina precedente. Alcuni numeri, detti numeri di Carmichael, in onore del matematico americano R. D. Carmichael che scoprì le loro proprietà, sono pseudoprimi in ogni base b. Il numero 561, prodotto di 3,11 e 17, è il più piccolo dei numeri di Carmichael; qui viene mostrato che è pseudoprimo nelle basi 2 e 3.

Supponiamo, per esempio, che n sia 257; allora $n - 1$ è 256, ossia 2^8, quindi tutti i fattori primi p di $n - 1$ sono uguali a 2. Per provare che n è primo, bisogna trovare un b tale che b^{256} sia congruo a 1 modulo 257, ma non è così per $b^{256/2}$. Sebbene il test non dia indicazioni su come trovare il particolare numero b, molti di tali numeri soddisfano le condizioni del teorema per qualsiasi primo n. Una ricerca casuale condurrà quasi sempre al successo. Quando b è uguale a 3, per esempio, 3^{256} è congruo a 1 modulo 257, mentre $3^{256/2}$ non è congruo a 256 modulo 257. Quindi 257 è primo. Sebbene sia anch'esso di tipo Monte Carlo, nel senso che b deve essere scelto a caso, una volta trovato il numero, il test di Lucas fornisce una rigorosa dimostrazione di primalità.

C'è un aspetto del test di Lucas che ne limita l'applicabilità a numeri di una certa forma: se non si possono trovare tutti i fattori primi p di $n - 1$, non può essere utilizzato. Ovviamente, se si pensa che sia primo, n è dispari e $n - 1$ è divisibile per 2. Ma questa buona partenza non basta; il test di Lucas non è generalmente applicabile agli altri $n - 1$ fattori, in modo semplice come nel mio esempio.

Se tutti i fattori primi di $n + 1$ possono essere trovati più facilmente di quelli di $n - 1$, esiste un altro test (anch'esso proposto per la prima volta da Lucas) per stabilire la primalità di n. Lehmer lo ha migliorato nel 1930 e per i numeri a cui può essere applicato in questa forma fornisce un procedimento che gira molto velocemente su un grande calcolatore. Grazie a esso è stata dimostrata la primalità dei più grandi numeri primi conosciuti, numeri della forma $2^p - 1$, dove p stesso è un numero primo. Tali numeri sono detti numeri di Mersenne, dal nome del matematico francese del XVII secolo che compilò per primo un elenco di numeri primi p per i quali riteneva che fosse primo anche $2^p - 1$. È evidente che se n è un numero di Mersenne, si conoscono tutti i fattori primi di $n + 1$: sono tutti uguali a 2.

Nel 1975 John Brillhart dell'Università dell'Arizona, Lehmer e Selfridge hanno mostrato come costruire un test di primalità per un numero n se si conoscono solo alcuni dei fattori primi di $n - 1$ o $n + 1$. Hugh C. Williams dell'Università di Manitoba ha portato questo tipo di test a una grande raffinatezza: per controllare la primalità di n sono sufficienti fattorizzazioni parziali di $n^2 + 1$, $n^2 - n + 1$ e $n^2 + n + 1$. Se però nessuno di questi si fattorizza con facilità, il test si arena. Per quanto questi metodi consentano di controllare molti numeri primi di 100 cifre, si è stimato che per condurre a termine l'esame di certi primi riottosi occorrerebbe un secolo di tempo macchina. Sarebbe quindi necessario un test di primalità che non dipendesse dalla particolare forma del numero in esame.

Nel 1980 Adleman e Robert S. Rumely dell'Università della Georgia hanno sviluppato un test che ha radicalmente modificato l'efficienza del controllo della primalità di grandi numeri che non abbiano una particolare forma. Nella formulazione originaria il test era in grado di saggiare la primalità di qualsiasi numero di 50-100 cifre in 4 - 12 ore di tempo macchina su un grande calcolatore. Henri Cohen dell'Università di Bordeaux e Hendrik W. Lenstra, Jr. dell'Università di Amsterdam hanno significativamente migliorato il test, tanto che attualmente gira 1000 volte più velocemente: un numero di 100 cifre può essere controllato in circa 40 secondi sul calcolatore Cyber 170-750 della Control Data Corporation.

Da dove proviene tanta efficienza del test di Adleman-Rumely? I suoi dettagli richiedono la conoscenza tecnica della teoria dei numeri algebrici, ma in sostanza è simile a quello di Fermat. Vengono costruiti due numeri ausiliari, detti il numero iniziale I e il numero euclideo E. Il numero I è il prodotto di svariati primi come $2 \times 5 \times 5 \times 7 = 210$. Il numero euclideo E deve invece questo nome al fatto che la sua definizione ricorda la dimostrazione di Euclide dell'esistenza di un'infinità di primi. E è il prodotto di tutti i primi p, q, r,... tali che i numeri $p - 1$, $q - 1$, $r - 1$,... sono tutti fattori di I. Il numero 70, per esempio, è un fattore di 210 e dato che 70 è minore di uno rispetto al primo 71, questo (71) è definito come un fattore di E. I fattori di 210 che sono predecessori di un primo sono 1, 2, 6, 10, 30, 42, 70 e 210 stesso. Quindi E è il prodotto dei primi 2, 3, 7, 11, 31, 43, 71, 211, ovvero è il numero 9 225 988 926. Il numero E deve essere costruito in modo da essere maggiore della radice quadrata del numero n di cui si controlla la primalità. Nel mio esempio con il numero iniziale 210 il metodo di Adleman-Rumely potrebbe lavorare con un n non maggiore di 10^{19}. Il tempo di calcolo è proporzionale a una potenza del numero I, pertanto questo andrebbe scelto il più piccolo possibile. Vi è una sorta di tensione dinamica fra I ed E. Perché il test sia valido, E deve essere grande; perché il test sia rapido, I deve essere piccolo. Inoltre, poiché E dipende da I, i numeri ausiliari non possono essere scelti indipendentemente uno dall'altro. Il numero 210 è un buon esempio di scelta di I, poiché si tratta di un numero relativamente piccolo con numerosi fattori che sono predecessori di un primo. Per dimostrare che il test di Adleman-Rumely è sempre veloce, era necessario verificare che fosse sempre possibile trovare dei numeri E e I opportuni e valutarne la grandezza.

Per un caso, si trattava di un lavoro già svolto. Nel 1955, Karl Prachar della Facoltà di agraria dell'Università di Vienna mostrò che esiste un'infinità di interi tra i cui fattori molti sono predecessori di un primo. Per applicare il risultato di Prachar al test di Adleman-Rumely nella sua forma originale, era necessario mostrare che il numero I può essere costruito in modo da essere «libero da quadrati», ossia non divisibile per alcun quadrato di un intero maggiore di 1. Cohen e Lenstra hanno recentemente dimostrato che, nella loro variante del test, la condizione della libertà da quadrati può essere eliminata. È stato possibile anche rafforzare il risultato di Prachar impiegando le scoperte di Patrick X. Gallagher della Columbia University e di Enrico Bombieri dell'Institute for Advanced Study e dell'Università di Pisa. Andrew M. Odlyzko dei Bell Laboratories e io abbiamo analizzato la costruzione dei numeri che possono recitare la parte di I nel nuovo test.

Dopo la costruzione di I ed E, sono stati sviluppati alcuni test analoghi a quello di Fermat per ogni coppia di primi p e q in cui p, il primo membro della coppia, è un fattore di E e q, secondo membro della coppia, è un fattore di $p - 1$. Il test non viene eseguito sui numeri interi, ma in rapporto ai cosiddetti interi algebrici che corrispondono a p e q. Un intero algebrico è un numero complesso che è radice di

I NUMERI PRIMI DI MERSENNE FINO A $2^{62\,982}$

VALORE DI p PER CUI $2^p - 1$ È PRIMO	$2^p - 1$	DATA DELLA DIMOSTRAZIONE	AUTORE DELLA DIMOSTRAZIONE	CALCOLATORE UTILIZZATO
2	3			
3	7	ANTICHITÀ	CITATO NEGLI *ELEMENTI* DI EUCLIDE	
5	31			
7	127			
13	8191	1461	CITATO NEL CODICE LAT. MONAC. 14 908	
17	131 071	1588	PIETRO ANTONIO CATALDI	
19	524 287			
31	2 147 483 647	1772	LEONHARD EULER	
61	19 CIFRE	1883	I. M. PERVOUCHINE	
89	27 CIFRE	1911	R. E. POWERS	
107	33 CIFRE	1914	R. E. POWERS, E. FAUQUEMBERGE	
127	39 CIFRE	1876 – 1914	ÉDOUARD LUCAS, E. FAUQUEMBERGE	
521	157 CIFRE			
607	183 CIFRE			
1279	386 CIFRE	1952	RAPHAEL M. ROBINSON	SWAC
2203	664 CIFRE			
2281	687 CIFRE			
3217	969 CIFRE	1957	HANS RIESEL	BESK
4253	1281 CIFRE	1961	ALEXANDER HURWITZ	IBM-7090
4423	1332 CIFRE			
9689	2917 CIFRE			
9941	2993 CIFRE	1963	DONALD B. GILLIES	ILLIAC-II
11 213	3376 CIFRE			
19 937	6002 CIFRE	1971	BRYANT TUCKERMAN	IBM 360/91
21 701	6533 CIFRE	1978	LAURA NICKEL, CURT NOLL	CDC-CYBER-174
23 209	6987 CIFRE	1979	CURT NOLL	CDC-CYBER-174
44 497	13 395 CIFRE	1979	HARRY L. NELSON, DAVID SLOWINSKI	CRAY-1

I numeri primi di Mersenne sono numeri primi predecessori di una potenza di 2. Il loro nome deriva da quello del matematico francese Marin Mersenne. Si tratta di numeri primi studiati fin dall'antichità, poiché, come mostrò già Euclide, se $2^p - 1$ è primo, $2^{p-1}(2^p - 1)$ è perfetto, ossia è uguale alla somma di tutti i suoi fattori, escluso se stesso. Se $2^p - 1$ è primo, anche p è primo, mentre l'inverso non è sempre valido. Non si sa se esista un'infinità di primi di Mersenne, né se esista un'infinità di composti di Mersenne, ma entrambe le asserzioni appaiono verosimili. Recentemente l'aumento dei numeri primi di Mersenne è andato di pari passo con la crescita delle capacità di calcolo. Secondo i dati di David Slowinski del Lawrence Livermore Laboratory, per dimostrare che $2^{8191} - 1$ non è primo, occorrono 100 ore macchina con un calcolatore ILLIAC-1, 5,2, ore con un IBM Sistema 7090, 49 minuti circa con un ILLIAC-II, 3,1 minuti con un IBM 360/91 e 10 secondi con il CRAY-1. Slowinski e Harry L. Nelson, anch'egli del Lawrence Livermore, hanno preso in esame tutti i numeri di Mersenne per valori di p fino a 50 000, senza trovare alcun primo maggiore di $2^{44\,497} - 1$. Recentemente Guy M. Haworth, Steven M. Holmes, David J. Hunt, Thomas W. Lake e Stewart F. Reddaway della International Computing Limited hanno proseguito la ricerca con un ICL-DAP, un supercalcolatore con 4096 elaboratori che lavorano in parallelo. Pur avendo portato il valore di p fino a 62 982, non hanno trovato alcun nuovo primo di Mersenne.

DIMENSIONE DEL NUMERO

TEST DI PRIMALITÀ	20 CIFRE	50 CIFRE	100 CIFRE	200 CIFRE	1000 CIFRE
DIVISIONE PER TENTATIVI	2 ORE	10^{11} ANNI	10^{36} ANNI	10^{86} ANNI	10^{486} ANNI
LUCAS, BRILLHART-LEHMER-SELFRIDGE, WILLIAMS	5 SECONDI	10 ORE	100 ANNI	10^{9} ANNI	10^{44} ANNI
ADLEMAN-RUMELY, COHEN-LENSTRA	10 SECONDI	15 SECONDI	40 SECONDI	10 MINUTI	1 SETTIMANA

Il tempo richiesto per il controllo della primalità di un numero varia notevolmente in base al tipo di test adottato. Nella tabella si assume l'uso di un calcolatore veloce. In particolare, per il metodo di divisione si assume l'uso di un calcolatore che esegua un milione di divisioni al secondo, indipendentemente dalla grandezza dei numeri esaminati. Il tempo di calcolo per una famiglia di test del tipo di quello ideato dal matematico francese Édouard A. Lucas è a volte poco confortante; alcuni numeri primi di forma particolare possono essere controllati molto più velocemente. La maggior parte dei tempi indicati sono risultati di stime, ma le caselle in colore riportano valori effettivamente ottenuti su calcolatori. Combinando i tre tipi di test, è possibile ottenerne uno ancora più veloce.

un'equazione algebrica a coefficienti interi, in cui il coefficiente della potenza di grado massimo dell'incognita è 1. Per esempio, $\sqrt{2}$, i (la radice quadrata immaginaria di -1) e $(-1 + i\sqrt{3})/2$ sono tutti interi algebrici in quanto radici delle equazioni algebriche $x^2 - 2 = 0$, $x^2 + 1 = 0$, $x^3 - 1 = 0$.

Se il numero n, di cui si controlla la primalità, non supera il test corrispondente a una delle coppie di primi p e q, allora viene classificato come composto, ma se li supera tutti non è certo che sia primo, anche se il numero di possibili fattori che resta da esaminare è piccolo. Adleman e Rumely hanno mostrato che ogni numero composto n che superi tutti i test del tipo Fermat deve avere i fattori primi in un insieme con esattamente I elementi. Lenstra ha mostrato che i numeri nell'insieme sono i residui delle potenze n, n^2, n^3... fino a n^I, modulo E. Provando con le divisioni se oltre a 1 e a n ciascuno dei numeri dell'insieme divide esattamente n, n è composto, altrimenti deve essere primo. Per quanto l'ultimo passo del procedimento di Adleman-Rumely assomigli a un processo di fattorizzazione, devo sottolineare che la sua conclusione risulta valida solo se n ha superato tutti i precedenti test di tipo Fermat. La maggior parte e forse tutti i numeri composti cadono per almeno uno dei test di Fermat, sicché non è necessario disporre di un fattore che debba essere trovato per divisione all'ultimo passo.

La velocità e l'applicabilità generale dei nuovi test hanno aperto la strada all'esame teorico di numeri precedentemente inaccessibili anche ai calcolatori più veloci. Supponiamo ora di sottoporre al test un numero con molto più di 100 cifre. In quanto tempo ci si può aspettare che il test emetta il suo responso? Questa e altre domande analoghe sono di grande importanza teorica per una branca della scienza dei calcolatori che prende il nome di teoria della complessità. Secondo una definizione correntemente accettata nella teoria della complessità, un test di primalità sarà «lento» dal punto di vista computazionale se non può essere eseguito in quello che è detto tempo polinomiale. Vale a dire, un test è lento a meno che il tempo impiegato per esaminare un numero n non sia minore di una potenza fissata k del numero di cifre di n. Indicherò il numero di cifre di n con $d(n)$. Si noti che $d(n)$ stesso è un numero, il cui numero di cifre sarà $d(d(n))$.

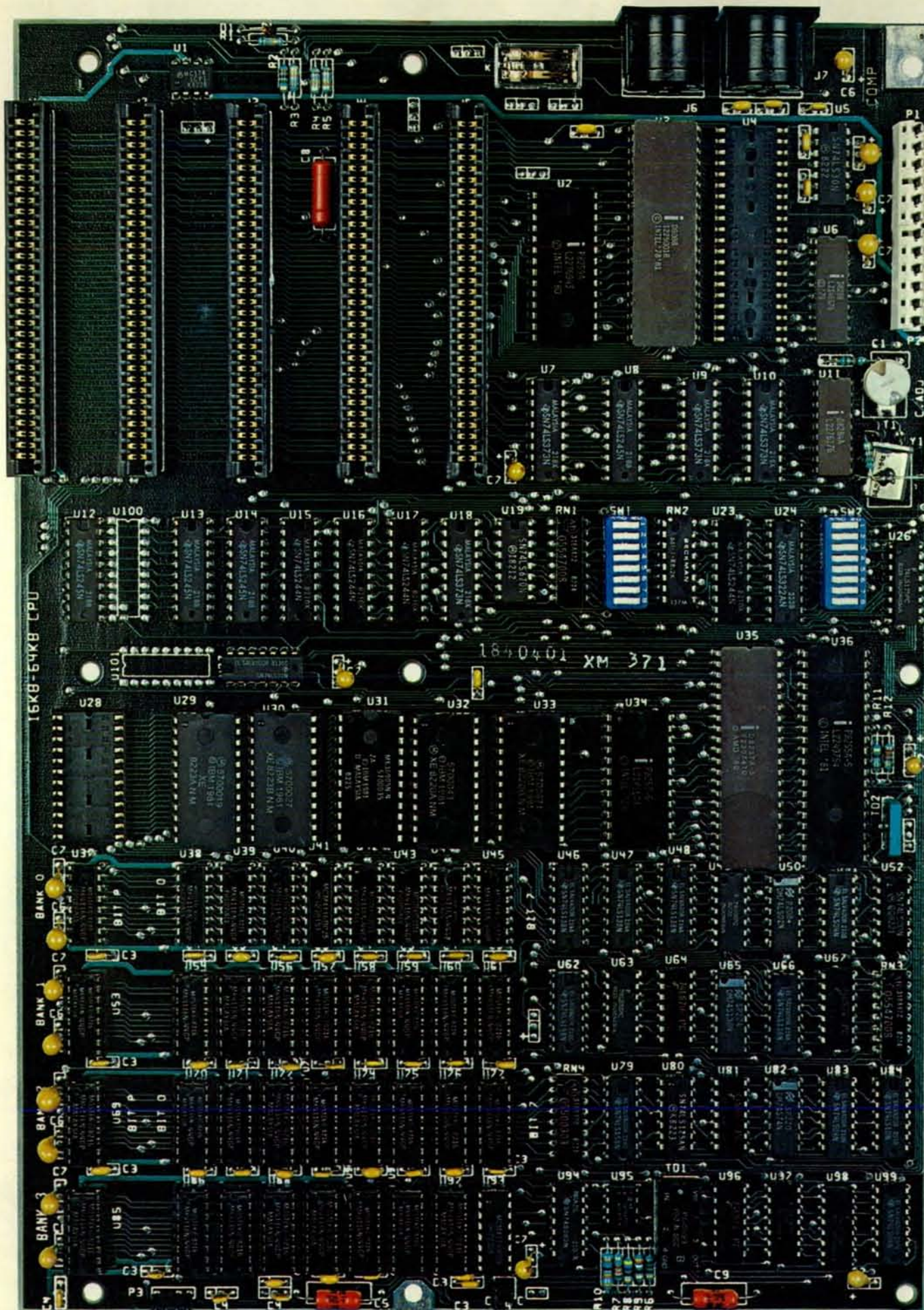
In base alla definizione fornita dalla teoria della complessità, i test di Adleman-Rumely e Cohen-Lenstra sono lenti. Il tempo di calcolo ha come limite $d(n)$ elevato alla $d(d(d(n)))$ moltiplicato per qualche costante c. L'espressione $d(d(d(n)))$ è il numero di cifre del numero di cifre del numero di cifre di n. Quale che sia la costante c per tale espressione, alla fine il prodotto di c per tale espressione eccederà la costante k, crescendo indefinitamente all'aumentare di n. Il limite del tempo di calcolo non è quindi polinomiale.

Tuttavia, per numeri relativamente «piccoli» il criterio proposto può essere fuorviante, in quanto $d(d(d(n)))$ in tal caso cresce all'infinito a una velocità decisamente ragionevole. Così, anche per un numero come 10^{1000}, $d(d(d(n)))$ è ancora uguale a 1. Il primo numero n per il quale l'espressione è uguale a 2 è $10^{999\,999\,999}$; in altre parole, per tutti i numeri inferiori il tempo di esecuzione dei nuovi test è limitato dalla potenza c di $d(n)$. Quindi i nuovi test non possono girare in tempo polinomiale, ma «quasi».

Ora che il test di primalità può essere eseguito rapidamente anche per numeri relativamente grandi, la nostra attenzione può rivolgersi al problema correlato della fattorizzazione. Dei progressi in questo campo avrebbero riflessi immediati sui sistemi crittografici basati sulla fattorizzazione. Gli sviluppi dei test di primalità non hanno direttamente a che fare con la fattorizzazione; d'altra parte, nessuno ha ancora dimostrato che questa sia intrattabile: nulla ci garantisce che domani qualcuno non possa inventare un metodo di fattorizzazione rivoluzionario. Una decisione circa la sicurezza a lunga scadenza dei sistemi crittografici a chiave pubblica basati sulla fattorizzazione coinvolge un giudizio soggettivo sui possibili progressi in materia. I recenti sviluppi nel campo dei test di primalità sottolineano la potenziale vulnerabilità di questi tipi di codice di fronte a nuovi risultati teorici.

Personal computer

Passando in rassegna hardware, software, applicazioni e diffusione di questi calcolatori, ci si rende conto che essi sono accessibili anche a persone non dotate di una preparazione tecnica specifica

di Hoo-min D. Toong e Amar Gupta

Se l'evoluzione dell'industria aeronautica negli ultimi 25 anni fosse stata spettacolare come quella dell'industria dei calcolatori, oggi un Boeing 767 costerebbe meno di un milione di lire e farebbe il giro del globo in 20 minuti, consumando una ventina di litri di combustibile. Prestazioni di questo genere rappresenterebbero grosso modo l'equivalente della riduzione di costi, dell'aumento di velocità operativa e della diminuzione di consumi energetici dei calcolatori. Il costo dei dispositivi logici per i calcolatori va diminuendo a un tasso del 25 per cento annuo e il costo delle memorie a un tasso del 40 per cento annuo. La velocità di calcolo è aumentata di un fattore 200 in 25 anni. Nello stesso arco di tempo il costo, il consumo di energia e le dimensioni dei calcolatori - a parità di potenza - sono diminuiti di un fattore 10 000.

INPUT/OUTPUT PER REGISTRATORE A CASSETTE
INPUT/OUTPUT PER TASTIERA
MICROELABORATORE INTEL
FESSURA PER ELABORATORE AUSILIARIO
FESSURE PER MODULI DI ESPANSIONE
CONDENSATORE DELL'OROLOGIO
OSCILLATORE DELL'OROLOGIO
INTERRUTTORI DI CONFIGURAZIONE DI SISTEMA
MEMORIA A SOLA LETTURA (ROM)
TEMPORIZZATORE
PORTA DI INPUT/OUTPUT
ZOCCOLINO DI ESPANSIONE
MEMORIA AD ACCESSO DIRETTO
MEMORIA AD ACCESSO CASUALE (RAM)
RITARDI, DECODIFICATORI E ALTRI CIRCUITI DI SUPPORTO
USCITA DELL'ALTOPARLANTE

Nella fotografia della pagina a fronte è mostrata la piastra principale del Personal Computer della IBM: i suoi elementi più importanti sono identificati nel disegno qui sopra. Le dimensioni della piastra sono di circa 25 × 36 centimetri. A essa sono fissati numerosi chip di silicio su cui sono integrati i circuiti; ciascun chip è un quadrato di meno di un centimetro di lato ed è inserito in un contenitore di plastica rettangolare, munito di elettrodi. I chip e altri elementi come resistori e condensatori sono collegati da conduttori stampati sulla piastra. Il microelaboratore, l'8088 a 16 bit prodotto dalla Intel Corporation, possiede 20 000 transistori e funziona a una frequenza di quasi cinque milioni di cicli al secondo. I «programmi di sistema» sono immagazzinati permanentemente nella memoria a sola lettura (ROM, *read-only memory*); nella memoria ad accesso casuale (RAM, *random-access memory*) vengono immagazzinati programmi e dati che cambiano di volta in volta.

Il risultato è la comparsa del calcolatore personale *(personal computer)*: con meno di un milione una persona può oggi avere a propria disposizione più o meno la stessa potenza fondamentale di calcolo di un grande calcolatore «*mainframe*» dei primi anni sessanta o di un minicalcolatore dei primi anni settanta. Vent'anni fa il costo di un calcolatore poteva essere giustificato solo se la macchina rispondeva alle esigenze di una grande organizzazione. I minicalcolatori introdotti agli inizi del decennio scorso si adattavano a una sezione o a un gruppo di lavoro all'interno della medesima organizzazione. Oggi il calcolatore personale può fungere da stazione di lavoro per una singola persona. Inoltre, così come è diventato possibile sul piano finanziario fornire un calcolatore al singolo lavoratore, gli sviluppi tecnici hanno reso sempre più «amichevole» l'interfaccia fra uomo e macchina, al punto che oggi molte funzioni di un calcolatore sono accessibili anche a chi non ha una preparazione tecnica specifica.

Il primo calcolatore personale fu posto in commercio nel 1975. Alla fine del 1982 i calcolatori personali funzionanti nei soli Stati Uniti erano, si calcola, più di un milione. Nel 1981 il totale del venduto, fra calcolatori personali e accessori, è stato, negli Stati Uniti, di 2,2 miliardi di dollari; secondo le previsioni nel 1985 si arriverà ai 6 miliardi di dollari. Sin dalla fine degli anni cinquanta, quando l'industria elettronica riuscì a stampare piccolissimi circuiti elettronici su *chip* di silicio, si è fatto un gran parlare di «rivoluzione del calcolatore» e «rivoluzione informatica». Finora quel che si è visto è stata un'evoluzione regolare, per quanto notevolmente veloce. Con il proliferare dei calcolatori personali, però, forse si è aperta la via per una vera rivoluzione nella gestione degli affari, nell'organizzazione delle attività personali e forse addirittura nel modo di pensare.

Anatomia di un calcolatore

Un calcolatore è sostanzialmente una macchina che riceve, immagazzina, manipola e comunica informazioni, suddivi-

L'hardware di un calcolatore personale comprende dispositivi per l'elaborazione e l'immagazzinamento di informazioni e per la comunicazione con l'utente e con altri dispositivi elettronici. Un insieme di conduttori paralleli, chiamato bus *(in colore)*, collega i componenti principali. L'unità di elaborazione, che in genere comprende non solo il chip microelaboratore stesso, ma anche vari altri chip ausiliari, esegue sostanzialmente tutti i calcoli e controlla tutto il sistema. Le informazioni possono essere immesse nel sistema attraverso una tastiera. Premendo un tasto si genera un segnale codificato, che è in corrispondenza biunivoca con quel tasto; il codice viene immagazzinato nella memoria del video e quindi appare sul visualizzatore, a tubo a raggi catodici. La memoria primaria, che è formata da chip di memoria a semincoduttori, contiene i programmi e i dati utilizzati in quel particolare momento: è una memoria ad accesso casuale, il che significa che il contenuto di ciascuna cella può essere esaminato o modificato indipendentemente da tutte le altre celle. La memoria di massa su disco in genere possiede una capacità maggiore della memoria primaria, ma è più lenta e le informazioni vengono recuperate a blocchi più voluminosi. Le interfacce collegano il calcolatore ad altri dispositivi, come una stampante o un modem (che consente l'accesso ad altri calcolatori attraverso la rete telefonica). In una interfaccia seriale le informazioni sono trasferite un bit alla volta; in una interfaccia parallela più conduttori trasportano più bit alla volta.

dendo ogni compito in operazioni logiche eseguibili su cifre binarie (stringhe di 0 e 1) ed effettuando tali operazioni a centinaia di migliaia o a milioni al secondo. Al cuore del calcolatore si trova l'unità centrale di elaborazione (CPU, *central processing unit*), che esegue le funzioni aritmetiche e logiche fondamentali e controlla il funzionamento di tutto il sistema. In un calcolatore personale la CPU è un microelaboratore: un unico circuito integrato su un chip di silicio di dimensioni, tipicamente, inferiori al centimetro per lato. Altri chip di silicio costituiscono la memoria primaria del calcolatore, dove possono essere immagazzinati istruzioni e dati. Altri chip ancora regolano l'ingresso *(input)* e l'uscita *(output)* dei dati ed eseguono operazioni di controllo. I chip sono montati su una pesante piastra per circuiti di materiale plastico: uno schema stampato di conduttori collega i vari chip e fornisce loro l'alimentazione. La piastra è chiusa in un contenitore; a volte le piastre sono più di una.

Le informazioni vengono immesse nel calcolatore mediante tastiera, oppure vi vengono trasferite da una memoria secondaria (nastri o dischi magnetici). Ciò che il calcolatore fornisce in uscita è visualizzato su uno schermo: il tubo a raggi catodici del calcolatore stesso (quello che si chiama usualmente *monitor*) o un comune televisore; può essere anche stampato su carta da una unità stampante separata. Si può collegare al calcolatore un dispositivo chiamato modem (modulatore-demodulatore) che consente la trasformazione dei segnali digitali in segnali analogici, che in questa forma possono venire trasmessi utilizzando una linea telefonica.

I chip, gli altri elementi elettronici e i vari dispositivi periferici costituiscono l'*hardware* del calcolatore. L'hardware non può fare nulla da solo; ha bisogno di una serie di programmi, ovvero di istruzioni: quello che si chiama collettivamente *software*. La parte fondamentale del software è un «sistema operativo» che controlla le attività del calcolatore e gestisce il flusso di informazioni. Il sistema operativo fa da tramite tra la macchina e l'operatore umano e tra la macchina e un programma «applicativo» che consente al calcolatore l'esecuzione di un'equazione differenziale, il calcolo di paghe e contributi o la stesura di una lettera. I programmi in genere sono immagazzinati in mezzi di memoria secondaria e vengono letti da questi e portati nella memoria primaria quando si rendono necessari per una particolare applicazione.

Il calcolatore personale

Un calcolatore personale è un piccolo calcolatore basato su un microelaboratore: è un microcalcolatore. Non tutti i microcalcolatori, però, sono calcolatori personali; ve ne sono infatti di dedicati a compiti specifici come controllare una macchina utensile o misurare l'afflusso di carburante nel motore di un'automobile; un microcalcolatore può essere un sistema di elaborazione e gestione testi, un videogioco o un «*pocket computer*», un calcolatore tascabile che non è esattamente un calcolatore. Un calcolatore personale è qualcosa di diverso: un calcolatore *stand-alone*, cioè autonomo e autosufficiente, che mette un'ampia gamma di capacità a disposizione di una singola persona. Noi definiamo calcolatore personale una macchina che possiede tutte le caratteristiche seguenti:

1. Il prezzo di un sistema completo è inferiore ai 5000 dollari. (Come equivalente italiano potremmo tracciare il limite superiore attorno ai 10 milioni di lire).

2. Il sistema incorpora (o può essere collegato a) una memoria secondaria costituita da unità a nastro in cassetta o a dischi.

3. Il microelaboratore può fornire una capacità di memoria primaria di 64 chilobyte (kB) o più. (Un chilobyte è pari a 2^{10}, cioè 1024 byte. Un byte è una stringa di otto bit, cioè cifre binarie. Un byte può rappresentare un carattere alfabetico oppure una o due cifre decimali. Una memoria da 64 kB può immagazzinare 65 536 caratteri, cioè un testo, tanto per avere un'idea, di lunghezza pari a un articolo medio-lungo di «Le Scienze».)

4. Il calcolatore può lavorare con almeno un linguaggio di alto livello, come Basic, Fortran o Cobol. In un linguaggio di questo tipo le istruzioni possono essere formulate a un livello piuttosto elevato di astrazione e senza dover tener conto delle operazioni particolareggiate dell'hardware.

5. Il sistema operativo rende possibile un dialogo interattivo; il calcolatore risponde immediatamente (o almeno velocemente) alle azioni e alle richieste dell'utente.

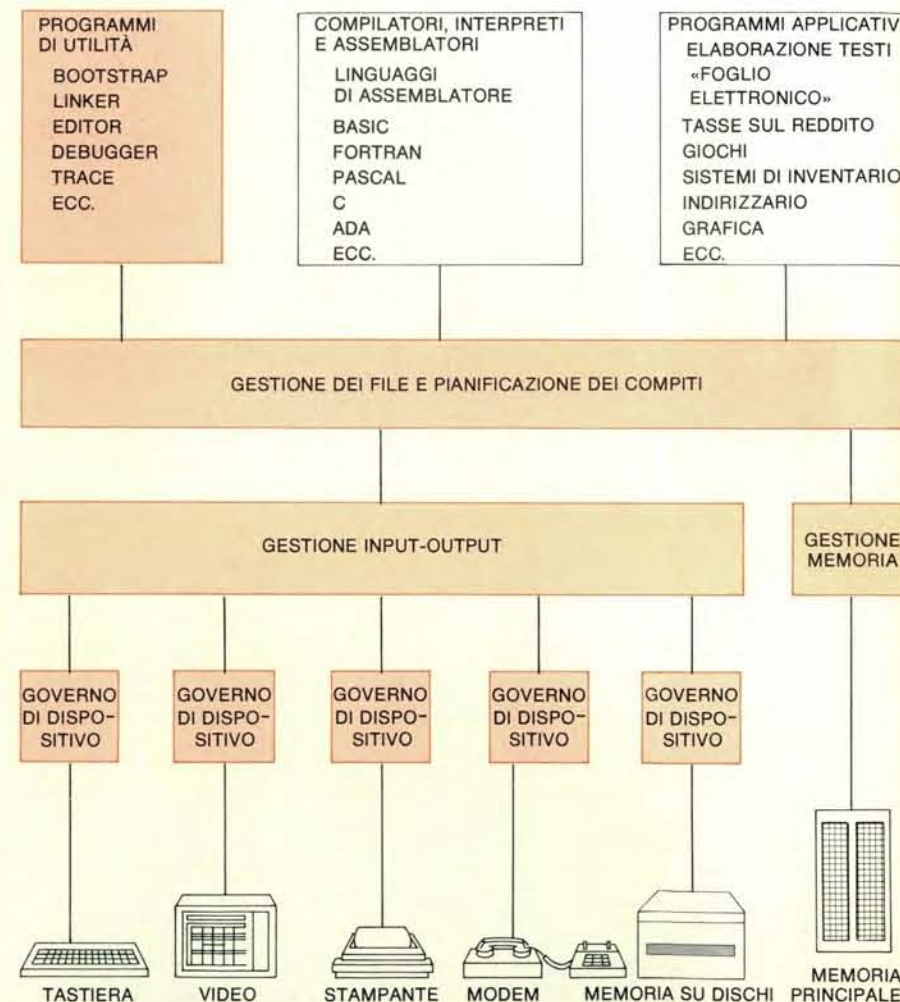
6. La distribuzione del prodotto avviene prevalentemente attraverso i canali distributivi tipici del mercato di massa e l'accento è posto sulla vendita a persone prive di una precedente esperienza su calcolatori.

7. Il sistema è abbastanza flessibile da accettare un'ampia gamma di programmi rivolti ad applicazioni diverse: non è cioè progettato per un unico scopo o per una sola, ben precisa, categoria di acquirenti.

La definizione sicuramente cambierà, non appena la tecnologia renderà possibili (e il mercato richiederà) l'inclusione di memorie di maggiore capacità e di hadware e software più specializzati nel sistema di base. Assolto, anche se per forza di cose un po' arbitrariamente, il compito di definire che cosa sia un calcolatore personale, passiamo ora a descriverne un po' più particolareggiatamente i componenti essenziali.

Microelaboratore e memoria

Due determinanti fondamentali della potenza di calcolo di un microelaboratore sono le sue dimensioni di «parola», che determinano la «ampiezza» del percorso dei dati del calcolatore, e la frequenza del suo orologio *(clock)* elettronico, che sincronizza le operazioni del calcolatore. La tendenza attuale è verso parole sempre più lunghe e frequenze sempre più elevate. Al crescere della lunghezza della parola, una operazione può essere completata in un minor numero di cicli di macchina; al crescere della frequenza, aumenta il numero dei cicli al secondo. In generale una parola più lunga consente anche l'accesso a un maggior volume di memoria. La prima generazione di veri calcolatori personali, entrati in commercio fra il 1977 e il 1981, possedeva microelaboratori a otto bit; i sistemi di più recente introduzione si basano su microelaboratori a 16 bit; nel frattempo sono stati costruiti anche chip microelaboratori a 32 bit, e presto anche questi verranno inclusi in sistemi di calcolatore completi. Oggi un microelaboratore a otto bit costa 5 dollari; un microelaboratore a 16 bit ne costa 50 e un microelaboratore a 32 bit ne costa 250. A mano a mano che la tecnologia migliorerà, i costi diminuiranno e proporzionalmente crescerà anche il numero dei calcolatori personali dotati di elaboratori a 32 bit. Fin verso il 1985, comunque, le dimensioni di parola saranno tipicamente di 16 bit. Per quel che riguarda invece la frequenza dell'orologio, si è passati dal megahertz (un milione di cicli al secondo) di pochi anni fa ai 10 megahertz e più di oggi.

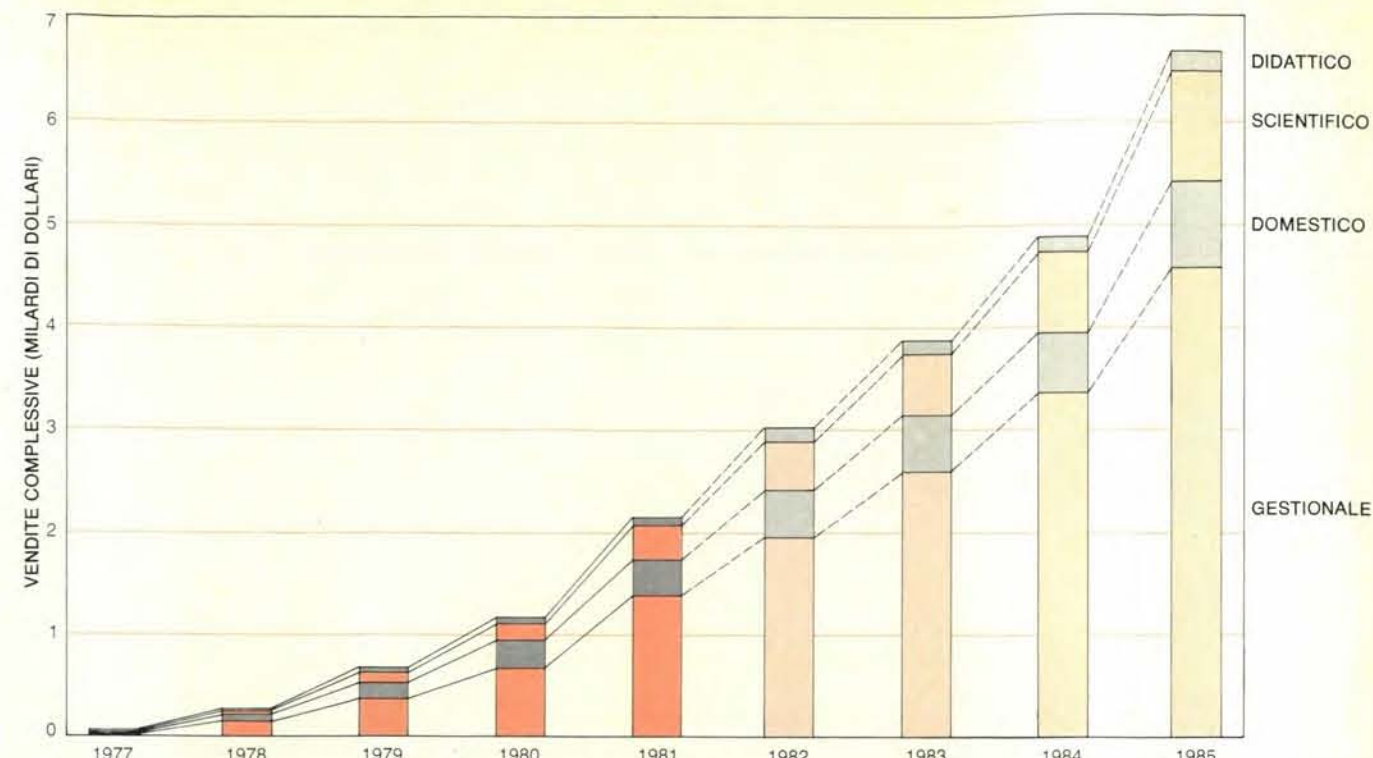
Esistono due tipi di memoria primaria: la memoria a sola lettura (ROM, *read-only memory*) e la memoria ad accesso casuale (RAM, *random-access memory*). Le memorie a sola lettura servono per informazioni che vengono «inscritte» nel sistema dalla ditta produttrice e che debbono rimanere immagazzinate in forma permanente: non possono venire modificate. Per un calcolatore dedicato a una specifica applicazione, come la gestione e l'elaborazione di testi *(word processing)* le informazioni su ROM possono comprendere il programma applicativo. Nel caso di un calcolatore versatile, non dedicato, fra le informazioni su ROM vi sarebbero sicuramente almeno i «programmi di sistema» più fondamentali,

Il software del calcolatore ha il suo centro nel sistema operativo *(in colore)*, un insieme di programmi che gestiscono le risorse del calcolatore, agiscono da supervisori nell'immagazzinamento di programmi e di altre informazioni e coordinano i compiti. I programmi applicativi sono quelli che eseguono qualche funzione sotto la direzione dell'utente. In teoria si potrebbero formulare programmi in grado di girare senza un sistema operativo, ma dovrebbero comprendere istruzioni dettagliate per l'assegnazione dello spazio di immagazzinamento sia nella memoria primaria, sia sui dischi, e per il funzionamento di tutti i dispositivi periferici. Tutti questi compiti sono assolti dal sistema operativo. Per poter essere eseguiti i programmi debbono essere in «linguaggio macchina» (in forma, cioè, di stringhe di cifre binarie): la necessaria traduzione è effettuata da programmi chiamati assemblatori, compilatori e interpreti. Assemblatori e compilatori traducono tutto un programma prima che questo giri; gli interpreti traducono ogni istruzione a turno, mentre il programma gira. Vari programmi «di utilità», considerati a volte parte del sistema operativo, possono aiutare l'utente nella stesura o nell'esecuzione di altri programmi. Un programma «di lancio» *(bootstrap)*, per esempio, fornisce le istruzioni iniziali quando si accende il calcolatore, e un programma di traccia consente di esaminare lo stato del sistema.

quelli che fanno partire il calcolatore quando si accende, interpretano il tasto premuto sulla tastiera o fanno sì che un «*file*» (ossia un gruppo di dati) memorizzato nel calcolatore venga stampato. Dato che il costo delle ROM sta calando, i costruttori preferiscono includere un numero maggiore di programmi di sistema su ROM, piuttosto che su mezzi di memoria secondaria.

La memoria ad accesso casuale è chiamata anche memoria di lettura/scrittura: vi si possono registrare nuove informazioni ed è possibile rileggerle ogniqualvolta lo si desideri. I chip di RAM immagazzinano informazioni che debbono essere periodicamente modificate (non importa che si tratti di programmi o di dati). Per esempio, si legge, da un disco di memoria secondaria, un programma per una particolare applicazione e lo si registra su RAM; una volta che il programma è in RAM, le sue istruzioni sono disponibili per il microelaboratore. Un chip di RAM stiva le informazioni in una schiera ripetitiva di «celle» microelettroniche, e ciascuna cella immagazzina un bit. La densità dei chip di memoria disponibili sul mercato (cioè il numero dei bit per chip) è aumentata di un fattore 64 nell'arco dell'ultimo decennio, con il risultato di ridurre di 50 volte il costo per bit. Cinque anni fa un singolo chip di RAM non poteva immagazzinare più di 16 chilobit (16 384 bit); oggi vari calcolatori personali dispongono di chip da 64 chilobit e si prevede che entro il 1984 i chip da 256 chilobit saranno largamente disponibili.

Il singolo chip di memoria è una schiera di bit, ma in generale le informazioni vengono trasferite dalla e verso la memoria primaria sotto forma di byte, e la capacità di memoria di un calcolatore si misura in byte. Tipicamente, nella configurazione base, un calcolatore personale dispone di una capacità di memoria RAM fra i 16 e i 64 chilobyte, capacità che può essere espansa con l'aggiunta di piastre o moduli di memoria addizionali. In generale è buona regola acquistare un sistema che possieda abbastanza memoria da immagazzinare i più lunghi fra i programmi applicativi che si prevede di dover eseguire. In generale i «*package*» di programmi standard già pronti specificano i requisiti minimi di memoria necessari.

Il mezzo standard per la memoria secondaria sono i *floppy disk*, dischetti flessibili in Mylar, in formato standard del diametro di 5 pollici e 1/4 o 8 pollici, rivestiti su una o ambedue le facce con un materiale magnetico. Le informazioni sono immagazzinate in tracce concentriche di piccolissime regioni magnetizzate; le variazioni nella direzione di magnetizzazione rappresentano gli 0 e gli 1 binari. Le informazioni vengono scritte su dischetto e recuperate dal dischetto grazie a una testina di registrazione che si sposta in senso radiale attraverso il disco in rotazione fino a una traccia particolare. La traccia a sua volta è divisa in più settori e di regola le informazioni vengono scritte o lette un settore alla volta. A seconda del particolare formato, vi sono da 8 a 26 settori per traccia e ogni settore può immagazzinare da 128 a 512 byte di dati. La capacità totale di immagazzinamento di un dischetto varia a seconda della densità di immagazzinamento dei dati lungo una traccia (che può raggiungere i 7000 bit per pollice), della densità delle tracce concentriche (si arriva a 150 tracce per pollice di raggio) e del numero di segmenti in cui è divisa ciascuna traccia. Per lo più i floppy disk hanno una capacità da 125 a 500 chilobyte, ma cominciano a essere commercializzati dischetti di densità superiore.

Una alternativa più costosa ai dischetti flessibili è rappresentata dai dischi Winchester, in cui il rivestimento magnetico è applicato a un piatto rigido in alluminio. Una unità a dischi Winchester per calcolatore personale può avere una capacità variabile da 5 a 50 megabyte (milioni di byte) e può trasferire i dati a una velocità molto superiore a quella di un floppy disk. Il disco Winchester, però, è sigillato permanentemente nella sua unità, mentre un floppy disk può essere tolto dal *drive* (ossia dall'unità dischi) e sostituito con un altro dischetto.

Le funzioni del sistema operativo sono illustrate dalla successione degli eventi necessari per caricare un programma applicativo. L'accensione del calcolatore *(1)* mette in funzione un programma di lancio che carica il sistema operativo nella memoria primaria. Il sistema operativo trasferisce un indice di *file* dalla memoria su disco alla memoria primaria; nell'indice di file è elencato l'indirizzo, cioè la posizione, di ogni programma e di ogni file di dati registrato sul disco stesso. In risposta all'istruzione successiva *(2)* il sistema operativo trova sul disco l'interprete Basic e, dopo essersi assicurato della presenza di uno spazio sufficiente, lo carica nella memoria primaria; quindi sul video appare un segnale che indica all'utente che l'interprete è pronto. (Alcuni calcolatori personali eseguono automaticamente il passo *2*, quando si accende la macchina.) Il sistema operativo viene richiamato per caricare il programma applicativo stesso *(3)*. A questo punto, con il controllo ancora all'interprete, può girare il programma applicativo. L'output sarà un nuovo file di dati nella memoria primaria, che poi potrà venire trasferito sul disco.

Si prevede che la crescita esponenziale del mercato dei calcolatori personali continui. Gli istogrammi danno il valore delle vendite di calcolatori personali per ciascun anno, relative al mercato statunitense per i quattro settori gestionale, domestico, scientifico e didattico.

Un mezzo di memoria secondaria più semplice e meno costoso è il nastro magnetico audio su cassetta. Una cassetta può immagazzinare una quantità di informazioni pari a quella memorizzabile su un dischetto flessibile di bassa capacità. Il tempo di accesso a un particolare indirizzo, cioè a una particolare locazione di memoria, è molto più lungo per il nastro che per il disco, sia perché la velocità del nastro è molto inferiore, sia perché su nastro le informazioni debbono essere disposte su un'unica successione lineare. Una caratteristica importante di tutti i mezzi di memoria secondaria di tipo magnetico è che le informazioni si conservano anche quando il calcolatore viene spento.

Uscita

Il mezzo principale di uscita per un calcolatore personale è un visualizzatore, di solito un tubo a raggi catodici: un monitor fornito insieme all'unità centrale, oppure lo schermo del normale televisore di casa. I visualizzatori a pannello piatto, che sfruttano la tecnologia dei cristalli liquidi o della scarica di gas cominciano a essere competitivi, in particolare per i piccoli sistemi portatili. Le immagini dei caratteri, necessarie per la visualizzazione dei testi, sono immagazzinate come figure di puntini in una speciale ROM, chiamata generatore di caratteri. La chiarezza del testo dipende dal numero di puntini utilizzati per la formazione di ciascun carattere. Un monitor tipico visualizza 24 linee di testo, ciascuna delle quali può contenere un massimo di 80 caratteri.

La visualizzazione di immagini grafiche, siano esse disegni ingegneristici, grafici o bersagli in movimento in un videogioco, richiede un software complesso e grandi quantità di memoria. Un disegno particolareggiato o una curva continua in un grafico richiedono una immagine ad alta risoluzione. La risoluzione è determinata dal numero dei pixel (elementi di immagine) indirizzabili dal calcolatore. Una immagine di 280 per 192 pixel in bianco e nero riempie più di 50 chilobit di RAM, mentre una immagine di 128 per 48 pixel richiede solo circa sei chilobit. Molti calcolatori personali possono produrre immagini a colori e, in questo modo, è possibile aumentare la quantità di memoria richiesta di un fattore quattro o più. Una immagine ad alta risoluzione, in particolare un'immagine a colori, può essere visualizzata chiaramente solo su un monitor.

Per molti scopi si può desiderare una copia stampata dell'uscita del calcolatore. Esistono molti tipi diversi di stampanti, che si differenziano largamente per il prezzo, la velocità di stampa e la qualità del testo stampato. Le stampanti termiche, che hanno un costo inferiore al milione (500 dollari), producono una immagine in una carta speciale alla velocità di circa 50 caratteri al secondo. Le stampanti a matrice di punti possono costare da meno di un milione fino a circa tre e possono essere molto veloci (fino a 200 caratteri al secondo). Il foglio viene «spazzato» da una matrice costituita da un minimo di 5 fino a 18 piccoli fili metallici: i segnali provenienti dal calcolatore guidano i fili contro un nastro inchiostrato, in modo da lasciare una figura di puntini sulla carta. La qualità dei caratteri che si vengono a formare dipende fortemente dalle dimensioni della matrice di punti disponibile per ciascun carattere; la matrice di punti è solitamente di 5 × 7 o di 7 × 9. Con opportuni programmi di controllo e con una capacità di memoria sufficiente la stampante a matrice di punti può generare immagini grafiche in bianco e nero o a colori.

Per lo più le stampanti termiche e a matrice di punti forniscono testi leggibili, ma che non si possono certo definire eleganti. Per avere una stampa «*letter quality*», cioè della qualità più adatta per una lettera, sono necessari dispositivi più costosi e più affini a una macchina per scrivere. Fra questi dispositivi rientrano le stampanti a margherita, il cui costo minimo si aggira sul milione e mezzo e che possono stampare fino a 55 caratteri al secondo. La testina di scrittura è un bottone rotante con 96 bracci radiali o più: su ciascun raggio vi è una lettera o un altro carattere. Mentre la margherita si sposta sul foglio, segnali provenienti dal calcolatore la fanno ruotare e mettono in moto

Negli Stati Uniti le vendite annue di calcolatori personali sono aumentate di 100 volte in sei anni, e più che raddoppiate nel 1982 *(in alto)*. Le ditte che hanno fatto da pionieri nella nuova industria sono state soppiantate da ditte che producono macchine adatte a un mercato più ampio *(in basso)*.

un martelletto che porta il raggio opportuno contro il nastro inchiostrato.

Il software

La capacità di memoria e di elaborazione delle informazioni di un calcolatore è determinata, fondamentalmente, dall'hardware, ma ben di rado l'utente ha occasione di maneggiarlo direttamente. Fra l'utente e l'hardware interviene una gerarchia di programmi, che considerati complessivamente costituiscono il software del calcolatore.

La parte del software associata più strettamente all'hardware è il sistema operativo. Per capire il tipo di compiti assolti dal sistema operativo, consideriamo la successione di passi per trasferire un file di dati dalla memoria primaria a un disco. In primo luogo è necessario assicurarsi che sul disco vi sia abbastanza spazio disponibile per accogliere tutto il file. Magari è necessario cancellare altri file per poter avere un numero sufficiente di settori vuoti contigui. Per il trasferimento stesso, bisogna richiamare dalla memoria primaria porzioni in successione del file e combinarle con informazioni di «economia domestica» che formino un blocco di dati tale da riempire esattamente un settore. Bisogna assegnare l'indirizzo di un settore a ciascun blocco, quindi bisogna trasmetterlo al disco. Bisogna calcolare dei numeri (chiamati somme di controllo) che consentono di identificare, e talvolta anche di correggere, eventuali errori nella memorizzazione o nella trasmissione. Infine, bisogna segnare da qualche parte dove è stato immagazzinato il file di informazioni.

Se tutti questi compiti dovessero venire svolti sotto la supervisione diretta dell'utente, il gioco dell'immagazzinamento di informazioni in un calcolatore non varrebbe la candela. In realtà tutto il procedimento può essere gestito dal sistema operativo; l'utente si limita a un'unica istruzione, come «Save file», cioè «registra il file». Quando si ha bisogno ancora delle informazioni in quel file, una istruzione analoga (del tipo, per esempio «Load file», cioè «carica il file») dà inizio a una successione di eventi in cui il sistema operativo recupera il file dal disco e lo riporta nella memoria primaria.

Nella maggior parte dei casi si scrive un programma applicativo perché venga usato con un particolare sistema operativo. Possono però esistere versioni di uno stesso sistema operativo per molti calcolatori diversi. Idealmente, allora, lo stesso programma applicativo potrebbe girare su vari calcolatori, purché tutti dispongano del medesimo sistema operativo; nella pratica, tuttavia, è necessaria qualche modificazione.

Il microelaboratore riconosce solamente un repertorio limitato di istruzioni, ciascuna delle quali deve essergli presentata sotto forma di una serie di cifre binarie. Una certa serie di cifre, per esempio, può dire all'elaboratore di caricare un certo valore della memoria primaria nel registro interno, chiamato accumulatore, e un'altra serie di cifre può dire alla macchina di sommare due numeri già presenti nell'accumulatore. È possibile scrivere un programma in questo «linguaggio macchina», ma il processo è noioso e porta facilmente a molti errori.

Il livello di astrazione immediatamente successivo è costituito da un «linguaggio assemblatore», in cui le serie di cifre binarie sono sostituite da simboli e parole più facili da ricordare. L'istruzione che dice di caricare nell'accumulatore può essere rappresentata, per esempio, da LOADA, e l'istruzione di sommare i contenuti dell'accumulatore può essere semplicemente ADD. Un programma chiamato assemblatore riconosce ciascuna di queste istruzioni mnemoniche e la traduce nella serie corrispondente di cifre binarie. In alcuni linguaggi assemblatori si può definire e chiamare con un nome una intera successione di istruzioni. Un programma scritto in linguaggio assemblatore, però, deve ancora specificare individualmente ciascuna operazione che l'elaboratore deve eseguire; il programmatore, inoltre, deve anche tener conto di dove sono memorizzati, nella macchina, ogni istruzione e ogni dato.

Un linguaggio di alto livello risparmia al programmatore il compito di adattare un procedimento all'insieme di istruzioni dell'elaboratore, tenendo conto in ogni momento di tutti i particolari della configurazione dell'hardware. Per sommare due grandezze, si può, con tutta semplicità, assegnare loro dei nomi, come X e Y. Invece di dire all'elaboratore dove può trovare, nella memoria primaria, i valori da sommare, il programmatore specifica l'operazione stessa, magari nella forma $X + Y$. Il programma, avendo tenuto una registrazione della locazione delle due variabili nominate, genera una successione di istruzione in linguaggio macchina, grazie alle quali i valori vengono caricati nell'accumulatore e sommati.

Esistono due grandi classi di programmi, chiamati rispettivamente «interpreti» e «compilatori», che traducono in linguaggio macchina un programma scritto in un linguaggio di alto livello. Un programma scritto in un linguaggio interpretato viene immagazzinato sotto forma di successione di istruzioni di alto livello. Quando si fa girare il programma, un secondo programma (l'interprete stesso) traduce ciascuna istruzione nell'opportuna successione di istruzioni in linguaggio macchina, che viene immediatamente eseguita. Con un compilatore tutta la traduzione viene completata prima di iniziare l'esecuzione. Il vantaggio presentato da un interprete è la possibilità di vedere singolarmente il risultato di ciascuna operazione. Un programma compilato, invece, in genere gira più velocemente, poiché la traduzione in linguaggio macchina è stata già effettuata.

Il Fortran è stato uno dei primi linguaggi di alto livello e oggi è disponibile in varie versioni (o «dialetti»). I programmi in Fortran sono compilati: le loro applicazioni principali sono in campo matematico e scientifico in generale. Il più utilizzato, fra i linguaggi di alto livello per i calcolatori personali, è il Basic, sviluppato negli anni sessanta da ricercatori del Dartmouth College. Secondo le intenzioni, il Basic doveva essere un linguaggio introduttivo per gli studenti di programmazione, ma oggi viene usato per applicazioni di ogni tipo. La maggior parte delle versioni del Basic sono interpretate. Esistono decine di altri linguaggi di alto livello che possono essere adottati da un microcalcolatore. La scelta di un linguaggio per un particolare programma spesso è basata sulla natura del problema in gioco; il Lisp, per esempio, è il linguaggio preferito da molti studiosi di intelligenza artificiale. Sono importanti anche considerazioni di stile nella programmazione personale: il Pascal ha acquistato larga diffusione, negli ultimi anni, perché si dice favorisca la stesura di programmi con una struttura di fondo chiara e facilmente comprensibile.

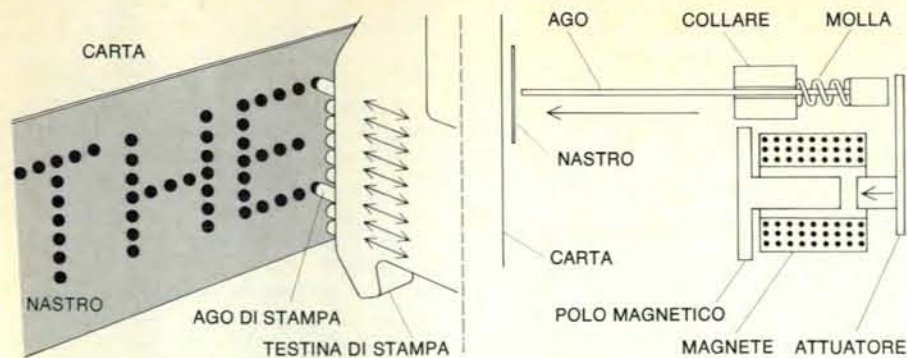
Programmi applicativi

I programmi applicativi sono quelli che, alla fin dei conti, determinano l'efficienza di un calcolatore nel soddisfare le esigenze dell'utente. Per questo è probabile che chi possiede un calcolatore personale finisca per investire più in software che in hardware. L'investimento può essere sotto forma di acquisto di programmi, o sotto forma del tempo (molto, in genere) che si deve spendere per la stesura in proprio dei programmi stessi. A meno che non ci si voglia dedicare a uno sforzo intenso di programmazione, l'ampiezza della base di software di un sistema (il numero delle applicazioni possibili) e la sua profondità (il numero dei diversi programmi disponibili per ciascuna applicazione) dovrebbero costituire elementi da prendere in seria considerazione nella scelta dell'hardware.

Si è sviluppata una fiorente industria di piccolissimi produttori di programmi applicativi, molti dei quali estremamente specializzati. Esistono, per esempio, programmi per la preparazione della denuncia delle tasse o (abbinati all'opportuna strumentazione di laboratorio) per analizzare migliaia di campioni di sangue all'ora, o ancora per progettare ponti. Altri programmi hanno applicazioni più ampie. Il software per il *word-processing*, la gestione e l'elaborazione di testi, è un esempio tipico: facilita la stesura e la redazione di documenti di ogni tipo, dalle lettere ai promemoria fino agli articoli di rivista come questo.

Il più famoso fra i programmi per calcolatori personali è chiamato VisiCalc ed è distribuito dalla VisiCorp. Viene definito un «tabellone elettronico». Il programma «stende» nella memoria del calcolatore e visualizza sul suo schermo una tabella con una larghezza di 63 colonne e un'altezza di 254 righe. L'utente può «scorrere» il tabellone verso destra e verso sinistra, oppure verso l'alto e verso il basso, in modo da portare in vista parti diverse. Ogni posizione (cioè ogni intersezione di una colonna e di una riga) sullo schermo corrisponde a un *record* (una unità di memorizzazione) nella memoria. L'utente definisce la propria matrice assegnando a ciascun record un'etichetta, un dato o una formula; la posizione corrispondente sullo schermo visualizza l'etichetta assegnata, il dato inserito o il risultato dell'applicazione della formula.

Facciamo un esempio molto semplice. Il cassiere di una società può inserire l'etichetta *Liquidi* nel record corrispondente alla colonna B, riga 1 (posizione $B1$), *Riserve* in $C1$ e *Totale* in $D1$. Può poi inserire 300 000 000 in $B2$, 500 000 000 in $C2$ e la formula $+ B2 + C2$ nella posizione $D2$. Sullo schermo, in $D2$ apparirà la scritta 800 000 000. Se il cassiere modifica il dato in $B2$ e scrive 200 000 000,

Il prezzo al dettaglio di un calcolatore personale riflette il costo per il produttore dei componenti dell'hardware, della manodopera e altri costi non relativi all'hardware, il profitto del produttore e la percentuale del distributore. In figura i costi si riferiscono a un calcolatore personale di costo e prestazioni elevati *(A)*, a un modello di medio livello *(B)* e a uno di basso costo e di modeste prestazioni *(C)*.

Una stampante a matrice di punti è relativamente poco costosa, veloce (fino a 200 caratteri al secondo) e flessibile: può produrre caratteri compressi, espansi o in neretto, o anche immagini grafiche, a seconda delle istruzioni che riceve dal calcolatore. La testina di stampa è una matrice verticale di aghi che vengono attivati selettivamente, mentre la testina scorre sulla carta, in modo da andare a premere un nastro inchiostrato contro la carta, formando di conseguenza una figura a puntini *(a sinistra)*. Qui ogni lettera maiuscola è data da un sottoinsieme di una matrice di sette punti in altezza e cinque in larghezza; sono disponibili altri due aghi per formare le discendenti delle lettere minuscole come la *p*. Gli aghi sono attivati da singoli solenoidi *(a destra)*. Il meccanismo illustrato in figura è quello della stampante a matrice di punti della Epson America, Inc.

il programma riduce il totale visualizzato in *D*2 a 700 000 000. Va detto, poi, che quel che viene inserito nei record *B*2 e *C*2 non è necessariamente un dato iniziale: può essere una funzione dei dati conservati in altri record.

L'industria

L'evoluzione del piccolo calcolatore personale è stata una conseguenza, forse inevitabile, della comparsa del microelaboratore. Fu nel 1971 che la Intel Corporation riuscì a incidere su un singolo chip tutti gli elementi di una unità centrale di elaborazione. Il primo microelaboratore era a quattro bit, ma nell'arco di un anno l'Intel riuscì a produrre un elaboratore a otto bit, di cui comparve poi nel 1974 una versione migliorata, l'Intel 8080. Non ci volle molto perché piccole ditte combinassero l'8080 con chip di memoria e altri componenti per produrre i primi microcalcolatori programmabili per controllo industriale e altre applicazioni specializzate. Nel 1975 la MITS, Inc., sviluppò il primo dispositivo abbastanza flessibile da poter essere considerato il primo calcolatore personale che sia stato messo in commercio: era l'Altair 8800 e il sistema base andò in vendita, diretto principalmente a chi aveva l'hobby del calcolatore, a 395 dollari in scatola di montaggio e a 621 dollari già montato. In quell'anno il meno costoso dei minicalcolatori era in vendita a 6000 dollari circa.

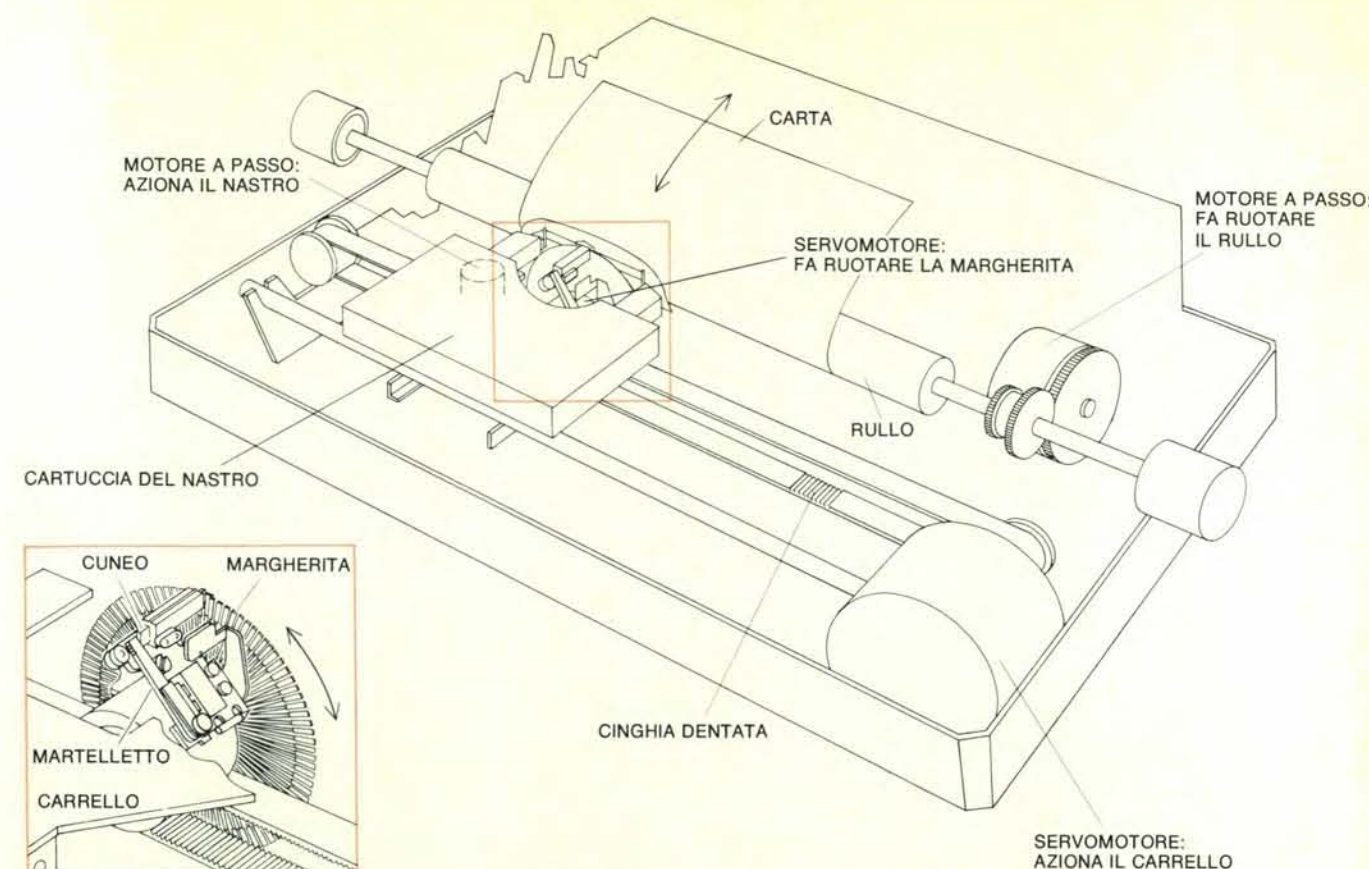
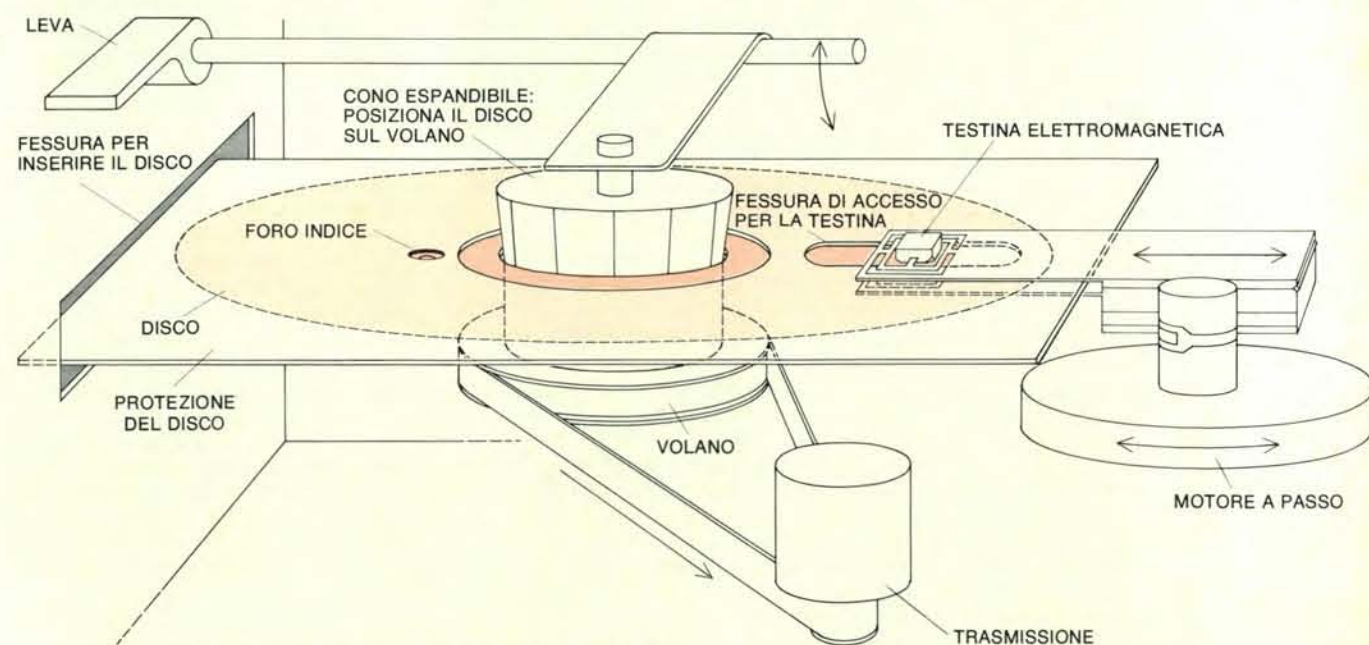
L'Altair non è più in produzione. Di fatto, vera ironia della sorte dell'industria dei calcolatori personali, le cui vendite annue sono aumentate di un fattore 100 nel giro di sei anni solamente, le ditte che hanno fatto da pionieri, come MITS, IMSAI Manufacturing Corporation e Processor Technology Corporation non sono riuscite a sopravvivere alla fase iniziale. I loro prodotti erano indirizzati principalmente agli appassionati, persone mosse da una profonda curiosità per i calcolatori e nella maggior parte dei casi fornite di conoscenze di elettronica, che desideravano - si potrebbe dire che addirittura bramavano - poter mettere le mani sull'hardware. Le ditte che hanno soppiantato i pionieri e che nel 1978 si erano accaparrate la fetta principale del mercato erano la Radio Shack, la Commodore Business Machines e la Apple Computer Inc. Queste società videro la possibilità di un mercato più ampio nel settore degli affari e in quello domestico: cominciarono a offrire sistemi «chiavi in mano», molto più accessibili anche a persone del tutto sprovviste di precedenti conoscenze nel campo dei calcolatori. Il successo di questa seconda generazione ha messo in allarme le grandi società produttrici di calcolatori *mainframe*, come la International Business Machines Corporation e la Burroghs Corporation e i produttori di minicalcolatori come la Digital Equipment Corporation e la Hewlett-Packard Company: i loro mercati tradizionali potevano venire erosi dai calcolatori personali, e così anche i grandi produttori tradizionali e consolidati nel campo dell'informatica sono scesi in lizza nel nuovo settore. E nuove ditte continuano a essere attratte verso questa attività.

Il mercato dei calcolatori personali può essere diviso in quattro aree: quella cosiddetta gestionale, quella domestica, quella scientifica e quella dell'istruzione. Il settore gestionale è diventato già, e di gran lunga, il maggiore. Nel 1981 rappresentava già 385 000 unità vendute (55 per cento del totale) con un valore al dettaglio di 1,4 miliardi di dollari (64 per cento del valore totale). Nei soli Stati Uniti esistono 14 milioni di aziende, e anche le più piccole sono potenziali acquirenti di un calcolatore personale. Cosa forse ancor più importante, nei soli Stati Uniti vi sono circa 36 milioni di «colletti bianchi» e molti, prima o poi, potrebbero finire per lavorare su un piccolo calcolatore di qualche tipo.

Attualmente il calcolatore personale si adatta particolarmente bene alle esigenze delle piccole aziende e dei liberi professionisti, come avvocati e medici. Le organizzazioni più grandi, tuttavia, stanno a poco a poco avvicinandosi all'idea delle stazioni di lavoro individuali, centrate sul calcolatore, che possono essere collegate l'una all'altra e a dispositivi centrali (grandi unità di memoria e stampanti, per esempio) attraverso reti locali (si veda l'articolo *La meccanizzazione del lavoro d'ufficio* di Vincent E. Giuliano, in «Le Scienze», n. 171, novembre 1982). I calcolatori personali sono già abbastanza potenti da poter svolgere la maggior parte dei compiti di una stazione di lavoro e si stanno sviluppando adatte reti locali. Nel 1985 in molte organizzazioni commerciali saranno sicuramente in funzione reti di calcolatori personali.

Il settore degli *home computer*, i calcolatori domestici, che è quello più visibile e su cui si riversa gran parte della pubblicità, nel 1981 ha rappresentato 175 000 unità vendute, per un valore di 350 milioni di dollari. Per lo più si tratta di sistemi venduti a fini ricreativi (principalmente per i videogiochi), ma queste unità possono essere utilmente impiegate come potenti sussidi didattici per i bambini, come sistemi per la gestione e l'elaborazione di testi, come centri di messaggi elettronici e come strumenti per la gestione delle finanze personali. Si sta sviluppando software che consentirà una gamma molto ampia di nuove applicazioni. Secondo le previsioni, il costo medio di un sistema domestico completo, che oggi si aggira sui 2000 dollari, è destinato a scendere a circa 1000 nel 1985 e a 750 nel 1990.

Le vendite nel settore scientifico sono state di 105 000 unità, nel 1981, per un valore di 336 milioni di dollari. I calcolatori rivolti alle applicazioni scientifiche e ad altre applicazioni tecniche sono tendenzialmente più potenti di altri calcolatori personali e posseggono in genere componenti che facilitano il collegamento a strumentazione per analisi e rilevamento. Il mercato è quindi caratterizzato da prodotti con hardware specializzato e una schiera di programmi altrettanto specializzati.

Il settore dell'istruzione è molto ampio, in potenza, ma dipende in misura critica dalla disponibilità di fondi, oggi piuttosto scarsa, per il sistema scolastico pubblico. Ciononostante nel 1981 le varie istituzioni scolastiche hanno acquistato 35 000 calcolatori personali per un valore complessivo di 98 milioni di dollari. L'istruzione assistita dal calcolatore (CAI, *computer-aided instruction*) coinvolge lo studente in una vivace interazione con la materia in quasi tutti i campi di studio e consente a ciascuno di procedere secondo il proprio ritmo di apprendimento. Si comincia a considerare la capacità di lavorare con un calcolatore come una necessità di base e presto in molti impieghi sarà richiesta addirittura qualche capacità di programmazione; chiaramente, il posto giusto per acquisire queste abilità è la scuola, primaria e secondaria. Conside-

Una stampante a «margherita» produce copie «qualità macchina per scrivere» a una velocità compresa fra 20 e 55 caratteri al secondo. Il disegno è una raffigurazione schematica di una stampante della Qume Corporation. La margherita di stampa ha un bulbo centrale in plastica sul quale sono disposte a raggiera 96 (su alcuni modelli 130) asticciole; all'estremità di ciascuno di questi raggi è fusa una lettera, un numero o un altro simbolo. In risposta a segnali del calcolatore, la margherita ruota in senso orario o antiorario per portare in posizione il simbolo giusto, quindi si ferma; il martelletto colpisce (con energia proporzionale alle dimensioni del simbolo: molto maggiore, per esempio, per una *W* che per un segno di virgola) il cuneo mobile, spingendolo contro l'estremità del raggio, che a sua volta va a premere il nastro inchiostrato sulla carta; il meccanismo di trascinamento e il nastro avanzano mentre la margherita ruota per portare in posizione il simbolo successivo.

Un sistema a *floppy disk* registra grandi quantità di informazioni su un dischetto flessibile in plastica rivestito di un materiale ferromagnetico. Il disco ruota a 300 giri al minuto in una custodia di plastica lubrificata. Una testina elettromagnetica si muove radialmente lungo la superficie del disco, grazie a un motore, fino a posizionarsi su una delle tracce concentriche in cui i dati sono immagazzinati sotto forma di una serie di inversioni nella direzione della magnetizzazione. La testina può leggere o scrivere: «sentire» le inversioni di magnetizzazione per recuperare informazioni o determinare una magnetizzazione per immagazzinare informazioni. Un foro indice, il cui passaggio è avvertito da un dispositivo fotoelettrico, sincronizza la registrazione o la lettura con la rotazione del disco. Il disegno è la rappresentazione schematica di una unità a dischi a doppia faccia prodotta dalla Qume. Vi sono due testine, che possono leggere e scrivere informazioni su ambedue le facce di un dischetto da 5 pollici e 1/4. Su ogni faccia del dischetto si possono registrare circa 160 chilobyte di informazioni, in 40 tracce concentriche.

rando che uno studente addestrato con il calcolatore di una particolare ditta un giorno o l'altro potrebbe facilmente diventare l'acquirente di un calcolatore di quella stessa ditta, la Commodore ha offerto a scuole e *college* i suoi calcolatori «tre al prezzo di due»; la Apple si è offerta di donare calcolatori personali a scuole primarie e secondarie statunitensi.

I produttori principali

Le industrie principali (in termini di stime di vendita per il 1982) sono Apple, Radio Shack, Commodore e IBM. Tutte stanno cercando di accaparrarsi la fetta più grossa del mercato nel settore gestionale, ma tutte cercano anche di seguire gli altri settori.

La ditta che conta sul maggior numero di vendite, non solo negli Stati Uniti, ma anche in tutto il mondo, è la Apple, il cui primo prototipo fu costruito nel 1976 in un garage. Nei primi quattro anni di esistenza la società è stata finanziata soprattutto da investimenti privati; è diventata società per azioni nel 1980, ma il 64 per cento del capitale è ancora in mano ai suoi promotori. Nel 1981 la Apple ha avuto un fatturato di 335 milioni di dollari (con un incremento del 290 per cento rispetto all'anno precedente) e profitti per 39,4 milioni di dollari (con un incremento del 340 per cento rispetto al 1980); detiene il 23 per cento del mercato statunitense e le sue vendite negli USA rappresentano solo il 76 per cento del totale. Gran parte del successo della Apple è da attribuire alla politica seguita da questa società, di incoraggiare i fornitori indipendenti di software a sviluppare e commercializzare prodotti compatibili con i calcolatori Apple. Per fare un esempio, sono disponibili per i calcolatori Apple 11 000 e più programmi applicativi, il 95 per cento dei quali è stato sviluppato da produttori indipendenti. Tutti e tre i modelli della Apple, attualmente in vendita, sono basati sul medesimo microelaboratore a otto bit.

La Radio Shack (dal 1963 proprietà della Tandy Corporation) prima di entrare sul mercato dei calcolatori era una ditta produttrice di apparecchiature elettroniche, con una rete di vendita al dettaglio: oggi i calcolatori rappresentano un quinto del suo volume d'affari. Il suo fatturato è cresciuto regolarmente, ma la sua «fetta» di mercato è scesa dal 50 per cento del 1978 al 22 per cento (secondo le stime) nel 1982. La Radio Shack dispone di una ampia gamma di prodotti, molti dei quali di propria fabbricazione, e di una distribuzione estremamente valida: oltre a 8000 punti di vendita per tutti i prodotti elettronici, dispone di una rete di centri per calcolatori, sia negli Stati Uniti che all'estero, i quali gestiscono vendite, leasing, servizi e addestramento. Il software è sviluppato sia internamente, sia da fornitori indipendenti.

La Commodore è una ditta canadese che ha iniziato la sua attività nel 1958 nel commercio di macchine per scrivere e nel 1976 ha acquistato la MOS Technology, la ditta produttrice del microelaboratore ancora utilizzato dai calcolatori Apple e Atari. Tra tutte le ditte, la Commodore è quella con il maggiore fatturato al di fuori degli Stati Uniti: attualmente detiene il 65 per cento del mercato europeo. Presenta un'ampia gamma di prodotti a basso costo (con un modello minimo in vendita a 150 dollari, intorno al mezzo milione di lire in Italia) e si è mossa bene nel settore dell'istruzione.

La IBM, massimo fornitore mondiale di apparecchiature per l'elaborazione di dati, domina da tempo il mercato dei *mainframe*, ma non ha avuto altrettanta fortuna nel campo dei calcolatori più piccoli, prima di entrare nel campo dei calcolatori personali, intorno alla metà del 1981. È riuscita a catturare una parte notevole del mercato (secondo le stime, il 14 per cento, nel 1982) in un tempo molto breve, seguendo la strategia di basarsi fortemente su fonti esterne non solo per il software, la distribuzione e l'assistenza, ma anche per l'hardware: il drive per i dischetti del calcolatore personale IBM è fornito dalla Tandon Corporation, il monitor arriva da Formosa e la stampante dal Giappone. La tastiera è fornita dalla IBM, e anche il nome è IBM. La IBM ha anche fondato una casa editrice che sollecita nuovi programmi di software da autori esterni.

Il successo della IBM ha conseguenze interessanti per il futuro del mercato dei calcolatori personali. Il settore è molto dinamico. Ditte statunitensi come la Xerox Corporation e la Atari, Inc., e vari produttori giapponesi (in particolare la Nippon Electric Co., Ltd.) sono in grado di superare le ditte che oggi guidano le vendite. Sono alle porte nuovi possibili concorrenti. Nel valutare le prospettive, ciascuno deve prendere in considerazione quali siano i requisiti per il successo commerciale. Quel che chiaramente non è indispensabile, a voler giudicare dalla strategia della IBM, è una capacità produttiva ben stabilita. Piuttosto sembrerebbe che i requisiti fondamentali siano la disponibilità di risorse finanziarie per l'acquisto dei componenti necessari e la capacità di commercializzare con successo un prodotto e di distribuirlo rapidamente su un'area molto vasta. Molte organizzazioni, tra cui alcune che oggi sono

Il software per la grafica gestionale aumenta nettamente l'utilità di un calcolatore personale ed è uno dei fattori che contribuiscono al crescente interesse suscitato da queste macchine. Il diagramma, una versione inglese dell'illustrazione di pagina 99, è stato generato su un calcolatore personale con Execu-Vision, un programma distribuito dalla Digital Systems Associates, ed è stato stampato con una stampante a matrice di punti. Il lavoro ha richiesto circa 15 minuti. La preparazione veloce di diagrammi di questo tipo, nonché di grafici, tabelle e anche disegni, con apparecchiature di basso costo, costituisce un ausilio molto valido nella stesura di relazioni.

Il calcolatore personale in Italia

Si racconta (in effetti questa «preistoria» ha già assunto, presso i cultori della materia, un'aura di leggenda) che il primo «personal» sia arrivato in Italia dall'America nel 1978, presentato allo SMAU, il salone dedicato all'ufficio che si svolge regolarmente nei padiglioni della Fiera di Milano, da una intraprendente ditta della capitale lombarda, che da allora è sempre stata molto attiva nella commercializzazione di macchine di questo tipo, per usi domestici e soprattutto per usi gestionali. Il primo importato era un Pet, prodotto dalla Commodore, una macchina assai curiosa a vedersi, nel modello di allora, con tastiera di formato non standard, video e CPU tutti integrati in un contenitore unico, di piccole dimensioni, con caratteri grafici e una ben precisa vocazione domestica, come peraltro il nome suggeriva (per chi non si fosse mai soffermato a rifletterci, «pet» in inglese significa animale domestico, in particolare di piccola taglia, cucciolo, e simili). In meno di cinque anni quella che sulle prime sembrava solo una curiosità è diventata un fenomeno di proporzioni assai notevoli anche nel nostro paese: pur non avendo ancora raggiunto le dimensioni e la diffusione che ha guadagnato negli Stati Uniti, il mercato dei personal e degli home computer può contare ormai anche in Italia centri di vendita appositi (le molte «computerie») e le prime catene (con grosse ambizioni, anche se nata in fondo da poco, quella organizzata dalla GBC, forte di una invidiabile esperienza nel campo elettrico ed elettronico) più o meno improntate a modelli statunitensi.

Sala di controllo dell'M20 (Ing. C. Olivetti & C., S.p.A., Ivrea).

Quasi tutte le marche presenti sul mercato d'oltreoceano sono state importate. È di questi giorni l'annuncio ufficiale da parte della IBM Italia della commercializzazione del loro Personal Computer *(si veda l'illustrazione alle pagine 96 e 97)* in Europa e in Italia in particolare; la distribuzione avverrà sia attraverso canali indiretti (concessionari autorizzati) sia attraverso i tradizionali canali diretti IBM. Il maggiore successo commerciale è toccato finora alla Apple, grazie anche a una efficiente distribuzione e a una pubblicità intensa ed efficace, seguita dalla Commodore, mentre le macchine della Radio Shack, a lungo dominatrici negli Stati Uniti, non hanno avuto un successo altrettanto ampio da noi: sicuramente a causa delle difficoltà di distribuzione (passata per varie mani nel giro di breve tempo) e anche per essere arrivate fin qui proprio nel momento in cui la parabola discendente delle vendite sul mercato americano era già iniziata.

Uno degli elementi che caratterizzano, in modo un po' curioso, il nostro mercato, è la fortuna che queste macchine hanno incontrato in campo gestionale (anche se non senza qualche delusione da parte degli acquirenti): Apple II, vari modelli Commodore e TRS 80 della Radio Shack vengono venduti in larga proporzione a utenti interessati ad automatizzare procedure di paghe e contributi, fatturazione e simili, attività per le quali peraltro queste macchine non sono state concepite originariamente. La Apple, per esempio, non ha mai sentito il bisogno di produrre, per il mercato statunitense, unità di memoria di massa di grande capacità per il modello II (unità a dischi rigidi, per esempio), mentre il distributore italiano si è recentemente preoccupato proprio di commercializzare unità di questo tipo, per soddisfare alle esigenze inevitabili di capacità di memoria piuttosto estese che sorgono nelle applicazioni gestionali.

Parallelamente all'introduzione di queste macchine, ha cominciato a espandersi anche l'editoria relativa: i volumi sull'hardware e il software dei personal computer formano già una nutrita biblioteca (ormai, tra l'altro, non più nemmeno solamente in traduzione, segno assai positivo della penetrazione del fenomeno) e alle riviste generali e a quelle più specificamente rivolte all'elaborazione personale, ma con spazio soprattutto per la valutazione delle macchine e le notizie di mercato, si sono cominciate ad affiancare una prima rivista esclusivamente dedicata al software per i personal computer («Personal software» delle edizioni Jackson) e una rivista esclusivamente rivolta agli utenti (e agli ammiratori) dei Commodore («Computer club»). I club di utenti che si scambiano informazioni e materiale relativi a una ben determinata macchina sono già numerosi, anche se non ancora ben organizzati: e a questa diffusione contribuiscono non poco i piccoli calcolatori, di uso decisamente domestico, che hanno più o meno ampiamente sfondato (verso il basso) il muro del mezzo milione di lire di costo (i Sinclair, il Vic 20 della Commodore, lo Home Computer della Texas Instruments).

Ma il panorama non è tutto di importazione; al contrario, almeno uno degli sviluppi più interessanti nel campo dei calcolatori personali porta un nome del tutto italiano, quello della Olivetti (senza voler togliere nulla alla qualità di altri costruttori nazionali di buon livello, dalla General Processor alla Saga, alla Kyber, che producono macchine che non hanno molto da invidiare a quelle d'oltreoceano). L'M20 della casa di Ivrea (sviluppato in effetti a Cupertino, nella «Silicon Valley» californiana dove hanno sede le maggiori ditte di calcolatori - non si dimentichi che la Olivetti è ormai, per quanto nostrana, una multinazionale) è un calcolatore personale della nuova generazione, come quelli della IBM, della Digital ecc., costruito attorno a un microelaboratore a 16 bit (lo Z8001) e con bus a 16 bit (il microelaboratore dell'IBM è un 16 bit, ma usa ancora un bus a 8 bit), dotato di un sistema operativo che la Olivetti ha voluto realizzare in proprio, con notevole impiego di energie (il PCOS, Professional Computer Operating System), senza ricorrere ad altri già esistenti. Pensato per il mercato internazionale, l'M20 ammette come linguaggi di programmazione una versione di Basic, Assembler e Pascal, ed è già stato affiancato da fratelli maggiori, M30 e M40, ancora più specificamente rivolti all'automazione del lavoro d'ufficio. È da notare, fra l'altro, che la casa di Ivrea, a differenza di altre, ha prestato particolare attenzione a fattori di carattere ergonomico (e le ricerche in questo senso procedono ancora) proprio in vista del fatto che la massiccia introduzione di calcolatori personali negli uffici e negli studi creerà problemi nuovi di sistemazione del posto di lavoro atta a garantire la confortevolezza e la salute dell'operatore.

È sicuramente presto per prevedere quale piega prenderà il mercato: un forte impulso potrebbe venire dalla diffusione dei servizi di telematica domestica (il servizio Videotel della SIP è entrato da qualche mese nella fase di sperimentazione sul territorio), ma non è detto che gli utenti possibili siano effettivamente già pronti per questi sviluppi; un altro punto di forza potrebbe essere la penetrazione nel mondo dell'istruzione, per la quale varie case si stanno preparando (ma le istituzioni scolastiche italiane sono notoriamente restie, in genere, al rinnovamento). Quel che è sicuro, invece, è che l'accento anche da noi già si sta cominciando a spostare dall'hardware verso il software: non serve a nulla avere una macchina potente ed efficiente, se non si dispone dei programmi per farla girare, e non molti sono in grado di realizzarli da sé. Sempre di più, il successo di una macchina tende a essere legato alla quantità di software disponibile: comincia l'età d'oro delle «software house», le ditte produttrici di software. Anche da noi cominciano a proliferare, ma... attenti alla qualità! *(V. Sala)*

attive in settori che non hanno nulla a che vedere con l'elettronica, hanno queste capacità e potrebbero facilmente acquisire tutti gli esperti di cui avessero bisogno. Organizzazioni così diverse fra loro come la CBS e la Coca Cola, Time-Life e la Prudential Insurance Co. possiedono le risorse finanziarie e quell'accesso ai mezzi di commercializzazione e di distribuzione che potrebbero consentire loro un rapido ingresso nel mercato dei calcolatori personali.

La distribuzione

I grandi calcolatori sono venduti direttamente dalla rete di vendita della casa produttrice, che tratta direttamente con la persona o l'azienda che ha intenzione di utilizzare il sistema. I margini di profitto sui calcolatori personali non sono abbastanza ampi da consentire un sistema di vendita diretta analogo. Per questo sono stati sviluppati molti altri canali di distribuzione, sia da parte dei produttori stessi, ma anche da parte di imprenditori dettaglianti.

I dettaglianti indipendenti che gestiscono un unico negozio si sono trovati in difficoltà con i calcolatori personali. Possono ordinare solo quantità molto limitate di un prodotto e in genere i loro capitali sono troppo esigui per potersi cimentare in una strategia di forte concorrenza. Sono stati soppiantati dalle grandi catene, come Computerland, negli USA, che nel 1981 ha venduto 200 milioni di dollari di calcolatori e accessori. Queste catene di negozi offrono i prodotti di diverse ditte; possono permettersi personale tecnico che consigli il cliente e possa fornire l'assistenza e la manutenzione sul lungo periodo. Catene di negozi meno specializzati, che vendono prodotti elettronici come impianti e componenti hi-fi hanno aggiunto al loro magazzino anche i calcolatori personali. Nei negozi di apparecchiature elettroniche la mancanza di competenza nel campo dei calcolatori ha costituito un handicap. L'affidabilità dello hardware, tuttavia, va migliorando e il software si fa sempre più standardizzato; con l'ingresso sul mercato delle ditte giapponesi (che, in particolare negli Stati Uniti, hanno forti collegamenti con le catene di negozi di apparecchiature elettroniche) è molto probabile che queste catene diventino uno dei canali principali di distribuzione.

I grandi magazzini in genere non hanno avuto molto successo nella vendita dei calcolatori personali. Da un lato, non possono assicurare una regolare assistenza; inoltre, uno studio ha messo in evidenza come chi acquista un calcolatore in media faccia quattro visite a un negozio, per un totale di sette ore, e i grandi magazzini non hanno una struttura adatta per questo genere di sforzo di vendita. I negozi di macchine per ufficio, invece, hanno molti contatti con le aziende locali e possono fornire la consulenza e l'assistenza tecnica necessarie. Negli Stati Uniti recentemente la Sears ha aperto nelle maggiori città punti vendita specializzati che trattano esclusivamente calcolatori personali, sistemi per la gestione e l'elaborazione di testi e apparecchiature ausiliarie.

I produttori stessi appoggiano numerosi canali di vendita per i loro prodotti. La Radio Shack dipende in gran parte dalla sua stessa catena di negozi al dettaglio. IBM, Xerox e Digital Equipment stanno aprendo propri punti vendita a complemento di altri canali di distribuzione. La Texas Instruments dispone di proprie esposizioni, dove il cliente può vedere i prodotti della ditta, scegliere e inoltrare un ordine che viene evaso da un magazzino centrale. I produttori possono anche trovare conveniente, a volte, un proprio personale di vendita diretta, per grandi commesse da parte di enti pubblici, grandi aziende e istituzioni accademiche, ma in questo modo corrono sempre il rischio di fare concorrenza ai dettaglianti che altrimenti potrebbero cercare di accaparrarsi vendite cospicue di questo genere.

Le reti di vendita per corrispondenza hanno avuto una presenza significativa nel campo dei calcolatori personali: trattano grandi quantità e possono offrire grandi sconti, ma non manutenzione sul posto né assistenza tecnica. I dettaglianti in grado di fornire una assistenza completa, poi, sono poco indotti a commercializzare un prodotto largamente disponibile a prezzi fortemente scontati.

Un nuovo tipo di sbocco, peculiare per il campo dei calcolatori personali, è la «società a valore aggiunto», che acquista hardware dal produttore, acquista o sviluppa periferiche e software per una applicazione specifica o per un tipo particolare di utente e offre un «pacchetto» completo. I servizi di una società del genere possono risultare particolarmente attraenti per le aziende che non dispongono di propri esperti o consulenti nel campo dei calcolatori.

A chi serve?

Nonostante le implicazioni del termine «personale» e nonostante l'immagine pubblicitaria tipica raffiguri i membri della famiglia raccolti attorno al calcolatore domestico per svolgere i compiti scolastici, far quadrare il bilancio di casa e abbattere invasori spaziali, è chiaro che per lo più i calcolatori personali sono acquistati da aziende o altre organizzazioni. Questo non rende necessariamente meno personale il calcolatore; la macchina può sempre essere dedicata a risolvere le esigenze di una sola persona. Negli Stati Uniti più di un quinto di tutti gli occupati lavorano in un ufficio e i costi degli uffici costituiscono più della metà dei costi totali di molte ditte, costi che vanno crescendo a un tasso superiore al 7 per cento annuo. I calcolatori personali possono aumentare la produttività dell'ufficio e dei «colletti bianchi». In una azienda che già possiede un grande calcolatore, i calcolatori personali possono alleggerire il carico sulle apparecchiature centrali, che possono così essere più ampiamente utilizzate per compiti di elaborazione di dati «a lotti», come la preparazione degli stipendi o i controlli d'inventario. I calcolatori personali rendono possibile la meccanizzazione di un'ampia gamma di lavori d'ufficio che fino a oggi sono stati svolti con macchine per scrivere, macchine calcolatrici e fotocopiatrici.

Si dice che i dirigenti dedichino più dell'80 per cento del loro tempo a prepararsi a incontri e «presentazioni» e a seguirli, a raccogliere informazioni o a prendere decisioni sulla base di una analisi delle alternative. I calcolatori personali hanno un'influenza su tutte e tre le attività. I nuovi programmi di *business graphics*, (grafica gestionale) consentono di produrre facilmente e rapidamente diapositive e materiale stampato per le riunioni. I dischi Winchester e i programmi per l'immagazzinamento e la gestione di grandi basi di dati consentono al singolo di esaminare una notevole quantità di informazioni, di individuare tendenze e identificare problemi. Programmi di manipolazione dei dati come VisiCalc consentono a un dirigente di valutare azioni alternative, consentono cioè di porre domande di quel tipo che si apre con «Che cosa succederebbe se...» e di avere una risposta quasi istantaneamente. In linea di principio compiti del genere possono essere eseguiti con un calcolatore *mainframe* centrale, ma è possibile assolverli meglio con un calcolatore personale, con un minore impegno di capitale e senza avere uno specifico addestramento tecnico.

Detto questo, resta il fatto che spesso non si può prevedere il ruolo preciso che un calcolatore personale deve svolgere all'interno di una azienda. Molte aziende hanno scoperto che, invece di soddisfare una esigenza nota, la presenza di un calcolatore personale spesso identifica un'esigenza precedentemente non individuata (più o meno come la presenza di un medico può portare alla luce un problema di salute in precedenza passato inosservato) e quindi riesce a soddisfarla.

Più difficile è dire se il singolo abbia bisogno del suo calcolatore personale, oppure possa trarne vantaggio o piacere. Per alcuni professionisti, certamente, la comodità di avere un calcolatore sempre pronto e a portata di mano è ben chiara. Altri possono acquistarne uno solo perché è disponibile e se lo possono permettere: le applicazioni verranno definite dopo. Le applicazioni specifiche nasceranno dalle capacità del calcolatore. Un calcolatore registra cose, può classificarle. Calcola. Può padroneggiare una gran massa di dati, modificare una variabile o più e vedere che cosa succede. Può far quadrare il conto in banca (o meglio, l'utente può far quadrare il proprio conto in banca con l'uso del calcolatore), può tener nota di appuntamenti o essere collegato a un impianto di allarme. Di per sé nessuna di queste applicazioni giustificherebbe l'acquisto di un calcolatore. Con curiosità e ingegnosità, però, chi possiede un calcolatore personale definirà le sue stesse applicazioni, strutturando il sistema a misura della sua personalità e dei suoi gusti.

TEMI METAMAGICI

di Douglas R. Hofstadter

«Presupposti riduttivi» e loro effetti sulla scrittura e sul pensiero

Questo articolo di Hofstadter ci ha creato non pochi problemi di traduzione, tanto da farci venire più volte la tentazione di lasciar perdere e, per un numero, «saltare» la rubrica. L'argomento, però, ci sembrava degno di nota e così, a conti fatti, siamo arrivati fino in fondo. Purtroppo l'italiano e l'inglese presentano più differenze di quelle apparenti a prima vista! Speriamo comunque che l'argomento interessi i nostri lettori - e che, soprattutto, i testi di Hofstadter non presentino più, in futuro, tutti questi problemi.

Un padre e un figlio stanno dirigendosi verso lo stadio quando la loro automobile rimane bloccata sulle rotaie della ferrovia. Si sente fischiare un treno in lontananza. Freneticamente il padre cerca di far ripartire la macchina, ma in preda al panico non riesce a girare la chiave e la macchina viene travolta dal treno. Arriva un'ambulanza e raccoglie gli infortunati. Durante il tragitto verso l'ospedale il padre muore. Il figlio è ancora vivo ma le sue condizioni sono critiche e richiedono un immediato intervento chirurgico. Appena giunto in ospedale viene trasportato in una sala operatoria d'emergenza e arriva un chirurgo che si aspetta un caso di routine. Alla vista del ragazzo, però, il chirurgo sbianca in volto e mormora: «Non posso operare questo ragazzo... è mio figlio.»

Come si risolve questo macabro indovinello? Come si può spiegare la cosa? Forse il chirurgo mente oppure si sbaglia? No. Lo spirito del padre morto si è in qualche modo reincarnato nel corpo del chirurgo? No. Forse il chirurgo è il vero padre del ragazzo mentre l'uomo morto era il padre adottivo? No. Qual è allora la spiegazione? Pensateci finché non ci arrivate da soli - insisto! State pur certi che saprete quando avrete trovato la risposta giusta.

Quando mi fu posto questo indovinello per la prima volta, qualche anno fa, trovai la risposta in circa un minuto. Eppure rimasi colpito dalla mia prestazione e da quella media del gruppo in cui mi trovavo, formato da uomini e donne colti e intelligenti. Non fui né il più veloce né il più lento. Un paio delle persone presenti rimasero più di cinque minuti a rompersi la testa prima di trovare la risposta e quando infine ci arrivarono rimasero a bocca aperta.

Comunque arriviamo alla risposta, rapidamente o con difficoltà, abbiamo tutti da imparare da questo indovinello: ci dà una penetrante visione del modo in cui i «presupposti riduttivi» permeano le nostre rappresentazioni mentali e incanalano i nostri pensieri. Un presupposto riduttivo è ciò che si considera vero in quello che si potrebbe definire il «più semplice» o «più naturale» o «più verosimile» modello possibile della situazione in esame. In questo caso il presupposto riduttivo è che il chirurgo sia un uomo. Il modo in cui vanno le cose nella nostra società oggi rende tale presupposto quanto mai plausibile. Ma il nodo centrale a proposito dei presupposti riduttivi è che essi sono automatici, non sono il risultato di un processo razionale di eliminazione. Voi non vi siete esplicitamente posti la domanda: a quale sesso è più plausibile che appartenga il chirurgo? È stata semplicemente la vostra esperienza passata a decidere quale sesso attribuire al chirurgo. I presupposti riduttivi sono per loro natura presupposti impliciti. Voi non eravate consapevoli di alcun presupposto sul sesso del chirurgo perché in tal caso l'indovinello non si sarebbe neppure posto.

Di solito è estremamente utile fidarsi dei presupposti riduttivi, senza i quali noi, o qualsiasi macchina cognitiva, non saremmo in grado di orizzontarci in questo mondo complesso. Non possiamo certamente permetterei di essere continuamente distratti da ogni tipo di eccezioni, teoricamente possibili ma inverosimili, alle regole generali o ai modelli che abbiamo costruito per induzione sulle molteplici esperienze passate. Dobbiamo azzardare ipotesi nel modo più sensato e lo facciamo in continuazione con grande abilità. Tutto il nostro pensiero è permeato di tali ipotesi sensate, presupposti di normalità, e si direbbe che tale strategia funzioni abbastanza bene. Per esempio, abbiamo buone ragioni per presupporre che i negozi lungo la strada principale di una città che attraversiamo non siano solo facciate di cartone. Probabilmente non avete il timore che la sedia su cui siete seduti possa sfasciarsi da un momento all'altro. Probabilmente l'ultima volta che avete usato una saliera non avete preso in considerazione l'ipotesi che potesse essere piena di zucchero. Senza molta difficoltà potreste trovare decine di presupposti a cui vi state affidando proprio in questo momento, ciascuno dei quali non è tanto *definitivamente* vero quanto *probabilmente* vero.

Questa capacità di ignorare ciò che è altamente improbabile, senza nemmeno prendere in considerazione se ignorarlo o meno. fa parte della nostra eredità evolutiva e discende dalla necessità di saper cogliere una situazione con rapidità ma anche con precisione: è una meravigliosa e raffinata caratteristica dei nostri processi mentali. Allo stesso tempo, però, questa meravigliosa capacità ci può sviare. Un esempio sono i presupposti riduttivi a proposito del sesso.

Quando scrissi il mio libro *Gödel, Escher, Bach: An Eternal Golden Braid,* impiegai la forma dialogica, una forma che mi paice molto. Ero così influenzato dal dialogo *Ciò che Tartaruga disse ad Achille,* di Lewis Carroll, che decisi di prendere in prestito i suoi due personaggi, i quali a poco a poco divennero i miei personaggi. Col procedere della stesura, mi trovai naturalmente portato a introdurre nuovi personaggi. Il primo fu Granchio; venne poi Formichiere. Bradipo e vari altri tipi pittoreschi. Come Tartaruga e Achille, i nuovi personaggi erano tutti maschi: il Sig. Granchio, il Sig. Bradipo e così via.

Questo avveniva agli inizi degli anni settanta ed ero perfettamente cosciente di ciò che stavo facendo. Per qualche ragione non riuscivo a inventare un personaggio femminile. Ero furente con me stesso, eppure non potevo fare a meno di pensare che introdurre un personaggio femminile «senza ragione» sarebbe stato artificioso e quindi fonte di distrazione. Non volevo mescolare la politica dei sessi, uno sgradevole problema del mondo reale, con gli eterei piaceri di un ideale mondo della fantasia.

Mi spremetti il cervello a lungo sull'argomento e arrivai anche a scrivere un complicato dialogo apologetico in cui discutevo con i miei personaggi la questione del sessismo nell'attività dello scrivere. Oltre ai miei amici Achille e Tartaruga, il cast comprendeva anche Dio come ospite d'onore. Pur nella sua ingenuità, si trattava di un onesto tentativo di affrontare alcuni problemi di coscienza che mi angustiavano. Il dialogo non ebbe mai una stesura definitiva e non fu inserito nel mio libro. Una serie di revisioni, però, lo trasformarono gradualmente nel «Ricercare a sei parti» che conclude il libro.

I miei rimorsi di coscienza mi portarono a introdurre qualche piccolo personaggio femminile: c'erano Prudenza e Imprudenza (che avevano una breve discussione sulla coerenza), Zia Hillary (una colonia cosciente di formiche) e tutti i membri dell'infinita serie Genio, Meta-genio, Meta-meta-genio e così via (*Genie, Meta-genie,* ecc., in inglese). Ero particolarmente fiero di questo tocco gentile, ma ciò non toglieva che i personaggi femminili avessero uno spazio molto limitato; non ne ero lieto, ma le cose stavano così.

Oltre ai dialoghi popolati di personaggi maschili, il libro era pieno di presupposti riduttivi di mascolinità, con una generalizzata scelta di pronomi maschili. Non avevo scuse per questo fatto; mi affidavo all'intelligenza del lettore, nella convinzione che capisse come spesso questi pronomi non contengano presupposti di genere ma si riferiscano semplicemente a una persona «unisex».

Col passare del tempo, però, sono giunto a una differente concezione del modo in cui il linguaggio scritto dovrebbe trattare persone di sesso non specificato, o persone che si suppongono specifiche ma scelte a caso. È un argomento sottile e non sostengo certo di avere delle risposte definitive; ho però elaborato alcune idee che mi piacciono e che possono forse essere utili anche ad altri.

Cosa mi diede la sveglia? Dato che ero già cosciente del problema, quale nuovo elemento mi portò a modificare le mie idee? Un incidente significativo fu la storiella del chirurgo. Rimasi sorpreso dalla mia reazione e da quella dei miei compagni. La maggior parte di noi aveva costruito ogni tipo di bizzarro mondo alternativo, invece di immaginarne uno in cui una donna potesse fare il chirurgo. Ridicolo! Questo fatto mi fece chiaramente capire quanto siano profondamente radicati i nostri presupposti riduttivi e quanto poco ne siamo coscienti. In potenza mi sembrava che la cosa potesse avere conseguenze molto maggiori di quanto si potrebbe ingenuamente pensare. Lungi da me ritenere che il linguaggio «faccia di noi ciò che vuole», che noi siamo i suoi schiavi, ma d'altra parte penso che dobbiamo fare del nostro meglio per liberare il linguaggio da usi che possono indurre o rafforzare presupposti riduttivi nella nostra mente.

Un episodio accaduto un paio d'anni dopo la pubblicazione del mio libro costituisce uno dei più chiari esempi di quanto ho detto. Parlando a un gruppo di persone dei dialoghi del libro, dissi di essere dispiaciuto del fatto che i personaggi fossero tutti maschi. Una donna mi chiese le ragioni della mia scelta e io risposi: «Beh, avevo cominciato con due maschi - Achille e Tartaruga - e sarebbe stato elemento di disturbo introdurre personaggi femminili senza apparente ragione se non la politica.» Ma appena mi sentii pronunciare queste parole, un orrendo pensiero si presentò per la prima volta alla mia mente: come facevo a sapere che Tartaruga, nel lavoro di Carroll, era davvero un maschio? Lo era sicuramente, vero? Mi sembrava di ricordarlo benissimo.

Eppure la domanda mi tormentava; avevo sotto mano una copia del dialogo di Carroll e la sfogliai per verificare. Grande fu il mio imbarazzo nello scoprire che in Carroll non c'è mai neanche un accenno al sesso di Tartaruga. Si trattava solo di un prodotto dei miei presupposti riduttivi?

Probabilmente no. Quando avevo sentito parlare per la prima volta del dialogo di Carroll, molti anni prima, era stato un maschio a descrivermelo. Molto probabilmente mi aveva passato il suo presupposto riduttivo. Mi potevo quindi proclamare innocente. Per di più, avevo letto alcune risposte al dialogo di Carroll su delle riviste di filosofia e anch'esse, come scoprii andandomele a rileggere, avevano immaginato una Tartaruga «sessuata», mentre Carroll aveva accuratamente schivato il problema. Pur sentendomi in qualche modo scagionato, rimanevo ugualmente turbato e continuai a chiedermi: cosa sarebbe successo se all'inizio avessi visualizzato una Tartaruga femmina? Cosa ne sarebbe stato di *Gödel, Escher, Bach?*

Una cosa che mi aveva dissuaso dall'usare personaggi femminili era l'abitudine che certi libri in lingua inglese hanno di usare i pronomi femminili per riferirsi al lettore o a una persona generica che venga brevemente menzionata. La cosa spiccava troppo e si finiva col mettere talmente in evidenza il problema del rapporto tra i sessi da far perdere spesso di vista il punto principale del passaggio. Mi sembrava una strategia troppo semplicistica che poteva finire col respingere molti lettori.

Eppure non potevo accettare il comportamento di certa gente, soprattutto ma non esclusivamente uomini, che rifiutavano di modificare le proprie abitudini linguistiche appellandosi alla «tradizione», alla «purezza linguistica», alla «bellezza della lingua» e così via.

Nell'introduzione a *Philosophical Explanations* di Robert Nozick, un mirabile e stimolante lavoro di filosofia, ho trovato la seguente nota: «Non conosco un modo di scrivere veramente neutrale per quel che riguarda il genere dei pronomi che non continui a distrarre l'attenzione, almeno quella del lettore contemporaneo, dal contenuto centrale della frase. Sto ancora cercando una soluzione soddisfacente.» Da questo punto in avanti, Nozick usa i pronomi maschili «*he*» e «*him*» quasi ovunque. La mia reazione fu di fastidio: Nozick si era davvero sforzato di trovare la soluzione? In parte il mio fastidio era indubbiamente dovuto ai miei sensi di colpa per non essere riuscito a far niente di meglio nel mio libro, ma in parte era dovuto alla sensazione che Nozick non fosse riuscito a vedere il fascino di una sfida su cui avrebbe potuto esercitare la sua capacità di intuito filosofico, dando così un contributo creativo alla società.

A quanto ricordo, il primo serio tentativo di «demascolinizzare» la mia prosa lo feci nel dialogo sul test di Turing uscito per questa rubrica nel luglio dell'anno scorso. Nello scrivere il dialogo, il sesso dei personaggi ondeggiava fluidamente nella mia mente: stavo infatti modellando i personaggi su un miscuglio di persone di mia conoscenza. Il personaggio più vicino a me come idee lo immaginavo sempre più come femmina che come maschio, e gli altri vacillavano.

Un giorno mi venne in mente di iniziare il dialogo presentando la questione di Turing «Si può in linea di principio distinguere, partendo da un dialogo scritto, una femmina da un maschio?» La questione si applicava così bene ai personaggi che ne discutevano che non riuscii a resistere alla tentazione di rendere qualche personaggio «ambisesso» - ambiguo in termini di sesso. Così chiamai uno dei personaggi «Pat». Presto mi resi conto che non c'era ragione di non estendere l'idea a *tutti* i personaggi del dialogo, che si sarebbe così trasformato in un vero indovinello per i lettori. Nacquero così «Sandy», «Chris» e «Pat».

Quel dialogo costituì per me una svolta decisiva. Anche se la sua totale uguaglianza sessuale era stata motivata dal desiderio di dare al dialogo un'interessante impronta autoreferenziale, mi scoprii sollevato per aver rotto lo stampo maschilista in cui mi sentivo prima rinchiuso e cominciai a cercare sempre nuovi modi per fare ammenda del mio passato sessismo. [Anche noi dobbiamo fare ammenda, almeno nei confronti dell'autore: nella traduzione del luglio 1981, l'ambisessismo di Hofstadter si è dileguato, e Pat è diventata una biologa, mentre Chris e Sandy sono inequivocabilmente uomini! *n.d.r.*].

Non era facile e continua a non esserlo. Per esempio, durante una lezione mi trovo a voler usare «she» («lei») per riferirmi a una persona non specificata, un biologo a caso, ad esempio, o un logico qualsiasi. Ma mi accorgo che non mi esce facilmente dalla bocca. Mi sono allenato abbastanza bene, invece, a evitare del tutto i pronomi di genere, «schivando» così il problema come fa Carroll. Ogni tanto dico «*he or she*» («lui o lei»), ma devo ammettere che spesso dico semplicemente «*they*». [Pronome inglese di terza persona plurale, che vale sia per il maschile che per il femminile. In italiano potremmo tradurre «loro», evitando la disambiguazione di «essi» o «esse». Ma «loro» non va bene in tutti i contesti, e comunque la necessità di concordare aggettivi e participi ci costringerebbe presto a smascherarci: «Loro sono andati» o «Loro sono andate», «Loro sono sicuri» o «Loro sono sicure»? Bisognerebbe usare solo aggettivi come «veloce» e «interessante», che non distinguono il genere, e usare solo forme verbali che non richiedono l'intervento del participio, come il passato remoto o l'imperfetto... *n.d.r.*]. Ovviamente, con l'uso di «*they*» (o di «loro») si casca dalla padella nella brace, in quanto si è semplicemente scambiata un'ambiguità maschio-femmina con un'ambiguità singolare-plurale. L'unico vantaggio, suppongo, è che non mi risulta ci sia/siano gruppo/gruppi attivamente in lotta per l'uguaglianza tra singolare e plurale. Una soluzione possibile è usare esclusivamente il plurale, riferendosi, ad esempio, a

«biologi» o «un'équipe di biologi», mai a «un biologo». In quel modo «they» («loro») si riferisce sempre legittimamente a un plurale. È una ben povera soluzione, però, in quanto è molto più efficace dare un'immagine di un individuo specifico. Non si può mai parlare di un corpo al plurale.

Un'altra soluzione, un po' più piacevole, che in qualche caso funziona bene, è quella di trasformare una situazione impersonale in una situazione più personale con l'uso del pronome «tu». In questo caso, tu che leggi o ascolti ti devi calare nella situazione e in qualche modo provare a sperimentarla in prima persona. [Che fatica tradurre questa frase! Hofstadter scrive, in inglese, «... your listeners or readers are encouraged...» e veniva subito spontaneo tradurre «voi ascoltatori o lettori siete incoraggiati...» *n.d.r.*].

A volte, però, la cosa si può ritorcere contro di voi. Supponiamo che stiate parlando degli strani effetti che le fluttuazioni statistiche possono produrre sulla vita quotidiana. Potreste scrivere qualcosa di questo genere: «Un giorno il vostro postino potrebbe avere talmente tanta posta da ritirare all'ufficio postale che ella potrebbe mettersi in cammino solo il pomeriggio». All'inizio la vostra avida lettrice Polly si costruisce un'immagine del suo amico postino, poi le viene detto che il postino è una donna. Il comprensibile stupore di Polly non è soltanto superficiale (per la collisione delle parole «postino» e «ella»); c'è un vero conflitto tra immagini, in quanto avete espressamente invitato Polly a pensare al *suo* postino, che guarda caso è un uomo. Polly si sarebbe stupita anche se aveste detto il «vostro portalettere». D'altra parte se aveste chiesto a Polly di pensare, ad esempio, al «portalettere di Henry», quell'«ella» non le avrebbe causato tanto stupore, e forse non gliene avrebbe causato affatto.

Quando insegno, cerco sempre di usare sostantivi neutri dal punto di vista del sesso come «portalettere» e «capo dipartimento», cercando di conseguenza di evitare pronomi di genere definito per riferirmi a quei sostantivi. [A noi le cose non vanno altrettanto bene, perché in italiano anche l'articolo distingue fra i generi e siamo costretti subito a dire *il* portalettere o *la* portalettere, anche se il sostantivo è indifferente al genere. Per questo, sicuramente, avrete trovato strano l'ultimo capoverso prima di questo, con tutta la discussione sul postino o il portalettere di Polly e Henry. *n.d.r.*]. Mi rendo conto, però, che si tratta di un'esibizione a mio esclusivo beneficio: non sarà certo evitando alcuni cattivi stereotipi che riuscirò a minarne la forza. Non credo proprio di riuscire a scuotere i miei studenti evitando di dire «egli», quando molti altri lo farebbero. Forse qualcuno può notare il mio «buon comportamento», ma si tratta di quelli che già sono in sintonia con questa problematica.

Allora perché non usare qua e là un inatteso «lei»? Non è la cosa più ovvia da fare? Forse; ma in molti casi, come sottolineava Nozick, la sua chiara motivazione politica finirà col distrarre più che con l'illuminare. Il problema è che, una volta usato un sostantivo come «portalettere» che si possa applicare a entrambi i sessi, la gente si costruirà un «nodo» mentale, una specie di gancio mentale a cui appendere varie qualità. (Se «nodo» non significa nulla per voi, immaginate un questionario con un certo numero di domande che richiedono risposte immediate.)

Ora, è ingenuo supporre che pochi secondi dopo la formazione di un nodo l'immagine sia, o sia mai stata, fluttuante in un limbo sessuale. È pressoché impossibile costruirsi un'immagine non effimera ed eterea di una persona senza assumere che sia una lei, o un lui. Dal momento in cui il nodo è costruito, se non risponderete alle sue domande sarà esso stesso a rispondere. (Si immagini che ogni spazio bianco del questionario sia riempito con una risposta riduttiva a matita, facilmente cancellabile ma da usarsi nel caso in cui non sia fornita alcun'altra risposta.) E sfortunatamente (la cosa vale anche per le più accese femministe) quelle inconsce risposte riduttive sono di solito sessiste. (Le donne possono essere sessiste tanto quanto un qualsiasi tizio.) Per esempio mi sono reso conto, con notevole disappunto, che i miei presupposti riduttivi hanno radici così profonde che quando *dico «letter carrier»* («portalettere»), e poi *«his or her route»* («la strada di lui o di lei»), spesso comunque *sto pensando «his route»* («la strada di lui»). [Qui in realtà, una volta tanto, l'italiano ci consente di aggirare l'ostacolo, in parte: noi diciamo comunque «la sua strada» o «il suo giro», e il genere del possessivo si riferisce al sostantivo che segue, non a «portalettere». L'inglese invece può specificare, con *«his»* o *«her»*, se il «suo» è «di lei» o «di lui». Noi ricorreremmo a queste ultime espressioni solo in caso di stretta necessità, perché in genere suonano un po' strane. Ma, davanti a una frase «unisex» come «chi è portalettere ha la sua borsa sempre a tracolla», pensereste a *una* postin*a*? *n.d.r.*]. La cosa è molto sconcertante, e mi rivela che, nonostante abbia dato buoni risultati a livello linguistico, il mio autoaddestramento non è ancora filtrato al livello delle immagini mentali.

Credo di aver trovato una soluzione di compromesso abbastanza appropriata che superi sia il metodo passivo, consistente nel limitarsi a evitare gli usi sessisti, sia quello attivo, consistente in eclatanti violazioni dello stereotipo. Invece di introdurre un genere non riduttivo *dopo* che la vostra lettrice ha costruito un'immagine riduttiva delle persone implicate nella situazione, non permettetele di impiantare le sue riduzioni. Rovesciate fin dall'inizio, in modo esplicito, i suoi (di lei) presupposti riduttivi.

Ho seguito questa via nell'apertura dell'articolo di agosto sui numeri troppo grandi, in cui raccontavo una vecchia storiella. Di solito il narratore inizia così: «Un professore stava tenendo una conferenza sul destino del sistema solare. Egli diceva...» Il professore è quasi sempre identificato in un maschio. Questo riflette forse le statistiche sul sesso degli astronomi, ma gli individui non sono le statistiche.

Come si potrebbe iniziare il racconto in modo migliore? C'è un intervallo, non lungo ma pur sempre un intervallo, tra il primo accenno al professore e la parola «egli». È un intervallo abbastanza lungo perché l'immagine riduttiva maschile si impianti solidamente, anche se implicitamente, nella mente dell'ascoltatrice. Facciamo allora in modo che la cosa non accada e trasformiamo fin dall'inizio il professore in una donna. Con questo non intendo certo che si debba iniziare il racconto con «Una professoressa stava tenendo una conferenza sul destino del sistema solare e...» Sarebbe orribile.

La mia soluzione fu quella di definire il sesso attraverso il nome. Inventai il nome pseudo-slavo «Professor Bignumska», la cui terminazione in «-a» indica che chi lo possiede è una donna. [Qui c'è anche un piccolo gioco di parole: *«big»* in inglese significa «grande» e *«num»* sono le prime tre lettere di *number*», cioè «numero». *n.d.r.*] A dire il vero, non tutti sono in sintonia con queste sottigliezze linguistiche e quindi alcuni si sorprenderanno nel leggere qualche riga dopo la frase «secondo i calcoli di lei fatti». Ma almeno lo stupore farà loro capire dove voglio arrivare.

La cosa peggiore è quando la gente, più che non afferrare la questione, la rifiuta in blocco. Nell'edizione francese di «Scientific American» («Pour la Science») il mio «Professor Bignumska» venne trasformato in «Monsieur le professeur Grannombersky». Non solo c'era stato un cambiamento di sesso ma chiaramente il traduttore si era accorto di ciò che intendevo fare e aveva deliberatamente cancellato ogni traccia indicativa trasformando in maschile la terminazione del nome. Provai un vero disappunto. Fui d'altra parte lieto di vedere che nell'edizione tedesca («Spektrum der Wissenschaft») rimaneva intatta l'appartenenza del professore al sesso femminile: era infatti chiamata «die namhafte Kosmogonin Grosszahlia». Non solo il cognome, ma anche il titolo aveva terminazione femminile. [Nella traduzione italiana è successo lo stesso: «la famosa professoressa Bignumska». Sull'edizione tedesca, come su quella francese, è stato reso anche il gioco di parole del cognome. *n.d.r.*]

Il fatto di riferirsi a membri di alcune professioni con sostantivi esplicitamente femminili e maschili crea certamente problemi. Cosa fate quando parlate di un gruppo misto di attori e attrici? Se non volete andar per le lunghe dovete necessariamente utilizzare il termine «attori». Perché una parola come «barista», con la sua terminazione assolutamente non indicativa, deve riferirsi a un maschio? Certamente i termini «oste» e «ostessa» non sono suoi sinonimi. D'altra parte, fa piacere vedere che «stewardess» e «steward» sono a poco a poco sostituiti dal termine generico di «assistente di volo».

Tutte le lingue che ho studiato sono in un modo o nell'altro afflitte da questo genere di problemi. Mentre in inglese abbiamo termini come *«poetess»* (poetessa) e *«aviatrix»* (aviatrice), i francesi, per riferirsi a una scrittrice o a una professoressa, devono usare *«une femme écrivain»* o *«une femme professeur»*: il riduttivo genere maschile è costruito nei nomi stessi. *«Ecrivain»* e *«professeur»*, cioè, sono entrambi nomi maschili; per renderli atti a riferirsi a donne bisogna utilizzarli come aggettivi che si legano (e modificano) alla parola *«femme»*.

Un'altra peculiarità del francese è l'espressione *«quelqu'un»*, che significa «qualcuno». Letteralmente significa «qualche uno» e, qualunque sia il referente, richiama il maschile «uno». Questo significa, per esempio, che se una sconosciuta bussa alla porta della casa di Nicole, la figlia di Nicole che è andata a vedere chi è dirà a Nicole: *«Maman, il y a quelqu'un à la porte»* («Mamma, c'è qualcuno alla porta»). È impossibile rendere al femminile il pronome: *«Maman, il y a quelqu'une à la porte»*. Sarebbe ancora più assurdo cercare di trasformare l'impersonale *«il y a»* («c'è») in una versione femminile, *«elle y a»*. Il maschile *«il»* è altrettanto impersonale dell'inglese *«it»* in *«It is two o'clock»* («sono le due»). Sicuramente nessuno suggerirebbe di dire *«They are two o'clock»*.

In inglese abbiamo alcuni fenomeni analoghi. Se una coppia di sconosciuti bussa alla porta di Paul, la figlia di Paul può dire: *«Daddy, someone's at the door»* («papà, c'è qualcuno alla porta»). Non dirà: *«Sometwo are at the door»*. Questo esempio illustra il fatto che il termine *«someone»* non veicola forti implicazioni di singolarità; si può applicare a un gruppo di persone senza sembrare strano. Forse, per analogia, *«quelqu'un»* non è così sessista a livello di immagine come il suo livello superficiale suggerirebbe. Ma questo è difficile da sapere.

Normalmente in francese, per parlare di un gruppo di persone misto o non specificato, si usa il pronome plurale maschile *«ils»*. Richiederà *«ils»* anche un gruppo i cui membri non sono stati determinati, ma che ha buone probabilità di includere almeno un maschio tra 20 femmine. Le donne, naturalmente, crescono con questo uso della lingua e lo seguono nello stesso modo naturale e inconscio dei maschi. Immaginate l'impressione che susciterebbe un serio tentativo di rovesciare questa antica convenzione? Come si sentirebbero gli uomini se il presupposto riduttivo fosse di dire *«elles»*? Come si sentirebbero le donne? Cosa sentirebbe la gente in generale se ci si riferisce con *«elles»* a un gruppo di persone formato da parecchi uomini e una donna?

Abbastanza curiosamente, ci sono circostanze in cui questo in un certo senso avviene. Esiste uno stile formale che si trova nei documenti legali o nei contratti in cui si usa la parola *«personnes»* per riferirsi a un gruppo di persone astratto o non specificato; il plurale femminile *«elles»* viene poi usato per riferirsi a quel nome. Dato che la parola *«personnes»* è femminile (si pensi al latino *persona*), questo è il pronome giusto da usare, anche se si sa che il gruppo a cui ci si riferisce è formato da soli maschi. [Lo stesso vale, ovviamente anche per l'italiano. *n.d.r.*]

Anche se questo uso è grammaticalmente corretto, se lo si prolunga nel corso del testo può dare al lettore una strana impressione, in quanto il nome originale è così distante da far apparire autonomo il pronome. Si ha l'impressione che il pronome dovrebbe a un certo punto trasformarsi in *«ils»*, e in effetti qualche volta accade. Quando non accade, il lettore può sentirsi a disagio. Forse questa è una mia reazione personale. Forse è semplicemente la reazione tipica di qualcuno abituato ad avere un pronome riduttivo maschile per un gruppo di persone non specificato.

Noi tutti, naturalmente, siamo membri di quel gruppo collettivo spesso chiamato «umanità», o semplicemente «uomo». Anche l'accesa femminista Ashley Montagu scrisse una volta un libro intitolato *Man: His First Two Million Years (Uomo: i suoi primi due milioni d'anni.* (Credo che sia stato molto tempo fa.) Molti sostengono che questo uso di «uomo» è completamente distinto dall'uso di «uomo» per riferirsi a individui e che è privo di implicazioni sessuali. David Moser ha intelligentemente rilevato la debolezza di questa affermazione. Egli osserva che nei libri si trovano frasi di questo genere: «L'uomo era per tradizione un cacciatore e teneva le sue femmine vicino al focolare dove potessero custodire i suoi figli». Non troverete mai frasi del tipo: «L'uomo è l'unico mammifero che non allatta sempre i suoi piccoli». Troverete piuttosto: «L'uomo è l'unico mammifero le cui femmine non allattano sempre i loro piccoli». E tanto basta per la «neutralità sessuale» del termine generico «uomo». Cominciai a prestare attenzione a queste anomali e presto mi imbattei, in un libro sulla sessualità, in questa gemma: «Non si sa in che modo l'Uomo usasse fare l'amore, quando egli era un selvaggio primitivo milioni di anni fa.»

Ma torniamo alle altre lingue. Quando passai alcuni mesi in Germania per lavorare alla mia tesi di dottorato, imparai che in tedesco il termine per «doctoral adviser» è «Doktorvater», letteralmente «dottore padre». Mi chiesi subito: e se il Doktorvater è una donna? È una «Doktormütter»? Quel titolo mi sembrava assurdo e pensai che una soluzione migliore fosse quella di aggiungere il suffisso femminile «-in», ottenendo così «Doktorvaterin», ossia «dottore padra». Mi sembra, però, che sarebbe preferibile un termine neutro.

L'italiano e il tedesco hanno una caratteristica comune: in entrambe le lingue il pronome di rispetto è derivato dal pronome femminile singolare, con l'unica differenza della maiuscola. In italiano è «Lei», in tedesco è «Sie». Ora, in tedesco il verbo associato ha la terminazione plurale, e quindi il rapporto con «lei» è in qualche modo diluito; in italiano, invece, il verbo rimane alla terza persona singolare. Per fare un complimento a un uomo potreste quindi dire: «Oh, come è bello Lei!» Naturalmente per un italiano potrebbe apparire altrettanto sconcertante il fatto che in inglese aggiungendo una «s» a un nome lo si rende plurale, mentre aggiungendo una «s» a un verbo lo si rende singolare.

Uno dei casi più strani è quello del cinese. Nel cinese Mandarino c'era per tradizione un solo pronome per «lui» e «lei», pronunciato «tā» e scritto

他

La parte sinistra di questo segno consiste del radicale di «persona», che indica che ci si riferisce a un essere umano di sesso non specificato. Curiosamente, però, nelle riforme linguistiche effettuate in Cina negli ultimi 70 anni, è stata introdotta una distinzione per cui ora ci sono forme scritte separate per il singolo suono «tā». Il vecchio segno è stato mantenuto, ma ora, in aggiunta al suo vecchio significato di «lei/lui», ha il nuovo significato di «lui» (l'avreste mai detto?) ed è stato inventato un nuovo segno per «lei». Il radicale del nuovo segno è lo stesso che per «donna» o «femmina» e il segno è

La nuova implicazione, non presente nel cinese prima di questo secolo, è che il tipo «standard» di essere umano è un uomo e le femmine devono essere indicate in modo particolare come «devianti». Rimane per me un mistero il motivo per cui non si è mantenuto il vecchio segno com'era, un pronome neutro, e non si sono semplicemente costruiti *due* nuovi segni, uno con il radicale femminile, come mostrato sopra, e uno con il radicale *maschile,* che avrebbe potuto essere qualcosa di questo genere:

(Questi tre caratteri cinesi sono stati creati su un calcolatore Vax con un programma per il disegno di caratteri scritto da David B. Leake e da me.) Per dare un esempio corrispondente (anche se esagerato) in inglese, si può immaginare una riforma politica in cui la parola *«person»* (persona) passasse a significare *«man»* (uomo) e per *«woman»* (donna) ci fosse detto di usare *«personess»*.

Il risultato è che nel cinese scritto non c'è più un pronome senza indicazione di genere. In precedenza, si poteva scrivere un'intera storia senza rivelare una sola volta il sesso dei personaggi: ora l'inten-

Achille e... Tartaruga (maschio o femmina?)

zione di essere ambiguo è essa stessa ambigua. Nel caso della storiella della cosmologa con la sua opzione riduttiva, è interessante considerare quale sarebbe la via migliore per le esigenze del femminismo. Sarebbe meglio che chi racconta la storia lasciasse non specificato per tutto il racconto il sesso del personaggio, appellandosi così alle opzioni riduttive della gente? Oppure sarebbe meglio che chi racconta si trovasse nella necessità di compromettersi?

Un vero e proprio tormento è per me l'uso corrente del termine popolare «*guys*». Spesso si sente descrivere un gruppo di persone con «*guys*», anche quando il gruppo comprende delle donne. In effetti, è abbastanza comune sentire delle donne che si rivolgono a un gruppo di altre donne con «*you guys*». La cosa mi sembra veramente strana. Alcune persone a cui ho sottoposto la questione, però, hanno fermamente sostenuto che quando «*guys*» è al plurale perde ogni traccia di maschilità. Stavo parlando dell'argomento con una donna ed ella sosteneva che «può aver mantenuto una connotazione maschile per *te*, ma non ce l'ha affatto per la maggior parte delle persone». Io non ero convinto, ma nulla di quanto dicevo riusciva a smuoverla dalla sua posizione. Alla fine fui fortunato perché in un ultimo tentativo di convincermi disse: «Ho perfino sentito usare *guys* per riferirsi a un gruppo di donne». Solo dopo averlo detto si accorse di aver contraddetto la sua stessa posizione.

Tali sono le sottigliezze della lingua. Spesso abbiamo semplicemente troppo poca coscienza del modo in cui lavora la nostra mente e di ciò che realmente pensiamo. È lì, perché lo percepiamo, ma troppo spesso la gente non ascolta se stessa; crede di poter conoscere se stessa senza ascoltarsi. Così mi è accaduto di sentirmi usare la prola «*chesspeople*» per riferirmi a quegli oggetti di legno che si spostano avanti e indietro sulla scacchiera, pur di non usare parole che terminino con «*man*» («*chessmen*» è il termine corretto per indicare i pezzi degli scacchi).

Il problema, per quanto riguarda i presupposti riduttivi, è che nella nostra società si manifestano dovunque. Li trovate in proverbi come «A ciascuno il suo», «Il tempo non aspetta nessuno» e così via. Li sentite quando i bambini (e gli adulti) parlano di scoiattoli e uccelli («Oh, guardalo come corre con quella ghianda in bocca!») Li vedete nei cartoni animati, in molti dei quali compare un povero zimbello, uno stupidone con cui «ogni uomo» può identificarsi, il cui destino è quello di essere schiacciato dal mondo intero, e noi ridiamo mentre subisce un colpo dopo l'altro. Perché non ci sono più spesso delle donne in questo ruolo?

Una sera, a casa di amici, stavo leggendo un delizioso libro per bambini intitolato *Frog and Toad Are Friends* (Rana e Rospo sono amici) e mi chiedevo perché Rana e Rospo dovessero essere entrambi maschi. Questo ci portò a discutere dell'argomento generale della rappresentazione femminile nella televisione e nel cinema per bambini. Parlavamo in particolare dei Muppets e ci meravigliavamo del fatto che tra i Muppets ci fossero così pochi personaggi femminili simpatici. Io sono un grande ammiratore di Ms. Piggy, ma penso ancora che ci sia qualcosa che non va se è l'unico importante personaggio femminile. Difficilmente si può considerare il suo un ruolo ideale.

Naturalmente, questo tipo generale di problemi non si limita a questioni di sesso. Si estende molto più in là, a gruppi di ogni tipo, grandi o piccoli. I fumetti di «The New Yorker», per esempio, pur essendo in un certo senso innocui, non fanno certo nulla per promuovere una modificazione nei presupposti riduttivi sui ruoli che le persone possono ricoprire. Quante volte si può vedere un dirigente nero o donna in questi fumetti (a meno, naturalmente, che ci siano per ragioni di battuta)? Lo stesso si potrebbe dire per la maggior parte degli spettacoli televisivi, la maggior parte dei libri, la maggior parte dei film... Non si sa proprio come combattere una struttura così monolitica.

C'è un ottimo e piacevole libro, che ho scoperto solo dopo aver quasi completato questo articolo, che potrebbe costituire un balzo da gigante per l'umanità nella direzione giusta. È *The Handbook of Nonsexist Writing*, di Casey Miller e Kate Swift (Barnes & Noble, 1980).

Uno dei più espressivi enunciati antisessisti che mi sia mai capitato di sentire faceva parte di un discorso pronunciato di recente a un banchetto di atleti di un college da Donald Kennedy, preside della Stanford University.

Kennedy ricordò che, 30 anni prima, anch'egli era un atleta di Harvard e aveva partecipato a un banchetto analogo. «Mi capita di chiedermi - rifletté assorto nei suoi ricordi - quale sarebbe stata la reazione se allora avessi predetto che ben presto... le donne avrebbero corso la maratona di Boston più velocemente di quanto avessero mai fatto gli uomini fino a quel momento. Da parte dei due terzi sarebbe venuta una risata incredula». Quindi precisò: «È proprio ciò che è successo. I miei compagni di corso sarebbero rimasti sorpresi di tale *evento* ma lo sarebbero rimasti ancora di più di fronte alla *tendenza generale*. Se guardiamo i migliori risultati mondiali della maratona maschile e femminile negli ultimi 10 anni, è chiaro che il record delle donne è cresciuto, nel decennio, a un ritmo sette volte superiore a quello del record degli uomini».

Il caso del nuoto è ancora più notevole. Kennedy ricordò che le squadre di Harvard e di Yale erano ai vertici della nazione nel nuoto ed entrambe arrivarono imbattute al loro tradizionale incontro di fine stagione. «Cosa sarebbe successo se in quella piscina fossero scese anche le attuali ragazze di Stanford?» chiese Kennedy. «*Umiliazione* è la parola giusta. Tanto per darvi un esempio, *sette* ragazze dello Stanford di oggi avrebbero battuto il mio amico Dave Heldberg, il grande stileliberista di Harvard, e *tutti* gli atleti di Yale, nei 100 metri. Le ragazze di Stanford avrebbero dominato i 200 dorso e rana, e vinto *tutte* le altre gare.

«Nella staffetta 400 stile libero ci sarebbe stato un distacco di 10 secondi tra Stanford e il primo degli uomini ad arrivare. Sapete quanto sono *lunghi* 10 secondi? Riuscite a immaginare la folla del Payne Whitney Gymnasium nel vedere una squadra di *ragazze* allineate contro le due migliori staffette di stile libero dell'Est e nel dover aspettare *così a lungo*, dopo l'arrivo delle ragazze, perché gli uomini portino a termine la loro gara?».

Il quadro dipinto da Kennedy era spiritoso, ma naturalmente quello che voleva dire era molto serio: «Io vi chiedo: se i luoghi comuni sulle capacità delle donne possono essere così brutalmente decimati in quello che è il più tradizionale campo di superiorità maschile, come possiamo mantenere le illusioni che abbiamo su di esse in altri campi?

«Qual è, in breve, la lezione da trarre dall'emergente eguaglianza *atletica* delle donne? Penso sia questa: coloro che assumono tutti gli altri presupposti, meno oggettivamente verificabili, sulle limitazioni delle donne, farebbero bene ad abbandonarli. Che sia frutto di malafede o ignoranza, il nonsenso è nonsenso. Ed è duro a morire».

Ecco un argomento su cui riflettere. Nel frattempo.

他